A FULLY ADAPTIVE ALGORITHM FOR PDE-CONSTRAINED OPTIMAL CONTROL WITH AND WITHOUT UNCERTAINTY, PART 1: AN INEXACT TRUST-REGION ALGORITHM AND THE DETERMINISTIC CASE*

SEBASTIAN GARREIS† AND MICHAEL ULBRICH†

Abstract. The purpose of this and a second paper is to develop and study an adaptive algorithm for the numerical solution of a class of optimal control problems, first for a deterministic semilinear state equation and then for a state equation with uncertain parameters. This paper investigates an adaptive inexact trust-region algorithm for minimizing a continuously differentiable function on a closed convex set in a Hilbert space. It allows for inexact objective function and gradient evaluations as well as an inexactly computed criticality measure. The global convergence of the method is proved and the implementation of the algorithm is discussed. The method is applied to the reduced version of a deterministic optimal control problem governed by a semilinear elliptic partial differential equation. A posteriori error estimators for the state and the adjoint equation are developed and it is shown how they can be used to control the adaptive discretization such that all error bounds required by the algorithm are satisfied. The investigations of this paper are carried out in a way that make them extendible to optimal control problems under uncertainty. Numerical experiments illustrate the adaptive behavior of the method.

Key words. optimal control, trust-region algorithms, adaptive algorithms, finite element discretization, a posteriori error estimation

AMS subject classifications. 35J61, 49J20, 49M15, 49M37, 65K05, 65M60, 90C30, 90C48

1. Introduction.

In this paper and its successor [12] we develop and analyze a class of adaptive optimization methods for solving deterministic and also *stochastic* semilinear optimal control problems which require only inexact function and gradient information. The—in both cases deterministic—control is subject to a general convex constraint. The current work and [12] span a rather wide range: In this paper, an adaptive inexact trust-region method is developed, its application to deterministic semilinear elliptic optimal control problems is studied, a posteriori error estimators are developed, and numerical experiments are conducted. The second part [12] builds on these results and extends them to semilinear elliptic optimal control problems under uncertainty, their discretization using low-rank tensors, develops suitable a posteriori error estimators for all sources of inexactness, and presents numerical results for the stochastic setting.

In this paper, we first develop and analyze an adaptive inexact trust-region method in a Hilbert space setting for the solution of optimization problems with a continuously differentiable objective function and a closed convex feasible set. It is a generalization of the algorithms in [15, 16] and can handle constraints by using an inexact projection onto the feasible set. We apply it to the reduced optimal control problem, which is analyzed in a suitable function space setting. The gradient of the reduced objective function is obtained by an adjoint approach. In each iteration of the trust-region method, the cost function, its gradient, and a criticality measure have to be evaluated up to a certain accuracy that is determined by the algorithm based on the current iterate. Errors in these quantities arise from the spatial—and in part 2 [12] also from the stochastic—discretization as well as from possibly inexact solvers for the discretized equations. Often, discretization errors can only be estimated up to unknown multiplicative constants. The proposed algorithm can deal with this challenge and

*Submitted to the editors DATE.

Funding: The work of the first author was partially supported by "TopMath", which is a graduate program of the Elite Network of Bavaria, Germany, and a graduate center of TUM Graduate School, and by the DFG/FWF International Research training Group (IGDK Munich – Graz) "Optimization and Analysis for Partial Differential Equations with Nonsmooth Structures."

†Technical University of Munich, Chair of Mathematical Optimization, Department of Mathematics, Boltzmannstr. 3, 85748 Garching b. München, Germany (garreis@ma.tum.de, mulbrich@ma.tum.de).

features global convergence to a first-order stationary point. We propose an error estimation framework which transforms the required error bounds to those for the computed state and adjoint state as well as the projection onto the feasible set. To fulfill these bounds we use an adaptive finite element (FE) grid refinement for the computation of the projection onto a box in $L^2(\Omega_u)$, where Ω_u is the control domain, and an adaptive discretization and solution technique for the state and adjoint equation, cf. [22].

In this paper, we present the essential ideas for a deterministic optimal control problem. In the second part [12], we extend them to a stochastic setting. There, many quantities and functions, such as the state $y \in Y$, are uncertain, i.e., they depend on a random parameter $\xi \in \Xi$, where $\Xi \subset \mathbb{R}^m$ is equipped with the probability measure $\mathbb{P}$. Still, all results derived in this part are valid for $\mathbb{P}$-almost every (a.e.) $\xi \in \Xi$ where then y corresponds to $y(\xi)$ and ξ occurs as a parameter in the PDE. Our results are developed in a form that allows to generalize them to the stochastic setting. For instance, the error estimates in section 4 are reused in [12, sec. 3]. The a posteriori error estimation technique in section 5 uses a reference operator and solves a corresponding reference equation. This will turn out to be an essential ingredient in the stochastic case, where the reference operator is deterministic and also serves as an efficient preconditioner for a low-rank tensor method.

A similar adaptive solution strategy is presented in [22, 23]. The main difference is that it uses an all-at-once approach, i.e., a trust-region SQP method computing a quasi-normal step towards feasibility and a tangential step towards optimality in each iteration. The approach of [22, 23] relies on the fact that the state space is a Hilbert space. The analysis of the stochastic model problem with a *semilinear* PDE in [12], however, makes it necessary to work with the Bochner space $L^p_{\mathbb{P}}(\Xi; H^1_0(\Omega))$ with $p > 2$, which is not a Hilbert space. Hence, we consider the reduced approach. Compared to earlier works, our approach is fully adaptive and reliable. While in [15] only the stochastic discretization is adapted, we additionally include the FE discretization and the algebraic errors into our adaptation strategy. Our earlier work [11] considered a fixed discretization and used a semismooth Newton method without addressing the question of global convergence. In the current paper, we ensure global convergence by a trust-region algorithm and use a semismooth Newton method as subproblem solver in our numerical implementation. We discuss more relevant literature about optimal control under uncertainty in the second part [12]. More references on multilevel approaches in the deterministic case can be found in [23, 22].

This paper is based on and uses parts of the dissertation [10]. It is organized as follows. The deterministic model problem is briefly discussed in section 2; the inexact trust-region algorithm is presented in section 3. Sections 4 and 5 deal with the error control procedure for the deterministic model problem and section 6 shows some numerical results including the adaptive generation of suitable FE grids.

2. Model problem. Let Y and U be Hilbert spaces and denote by Y^* the dual space of Y. We consider a state equation of the form

$$E(y,u) = Ay + N(y) - Bu - b = 0 \in Y^* \tag{2.1}$$

with a bounded linear operator $A : Y \to Y^*$ that is strongly monotone with constant $\underline{\kappa} > 0$ (in the sense of [21, Def. 25.2]), a monotone, twice continuously differentiable operator $N : Y \to Y^*$, $B \in \mathscr{L}(U, Y^*)$, and $b \in Y^*$. Furthermore, $y \in Y$ is the state and $u \in U$ is the control. The objective function

$$J(y,u) = \tfrac{1}{2}\|Qy - \hat{q}\|_H^2 + \tfrac{\gamma}{2}\|u\|_U^2 \tag{2.2}$$

is of tracking type with a real Hilbert space H, a desired state $\hat{q} \in H$, $Q \in \mathscr{L}(Y,H)$, and $\gamma > 0$. With a nonempty, closed, and convex set of admissible controls $U_{\mathrm{ad}} \subset U$, we consider the

following optimal control problem:

$$\min_{y\in Y,u\in U} J(y,u) \quad \text{s.t.} \quad E(y,u)=0, \quad u\in U_{\mathrm{ad}}. \tag{2.3}$$

2.1. A class of semilinear, elliptic PDEs. We now make the state equation more concrete and discuss a class of elliptic PDEs. Let $\Omega\subset\mathbb{R}^n$, $n\in\{2,3\}$, be an open, bounded domain with Lipschitz boundary $\partial\Omega$. We choose $Y:=H_0^1(\Omega)$ as state space, and equip it with the inner product $(v,\tilde v)_{H_0^1(\Omega)}=\int_\Omega \nabla v\cdot\nabla\tilde v\,\mathrm{d}x$. The control space is $U:=L^2(\Omega_u)$, where Ω_u can, e.g., be a measurable subset of Ω for example or—for finite-dimensional controls—$\Omega_u=\{1,\dots,n_u\}$ for some $n_u\in\mathbb{N}$. The control acts on the system via a bounded linear operator $D\in\mathscr{L}(L^2(\Omega_u),L^2(\Omega))$. Boundary control with, e.g., Ω_u being a subset of $\partial\Omega$ can also be handled by the algorithm presented later, but is not discussed here. We have a right-hand side term $f\in L^2(\Omega)$ and a coefficient function $\kappa\in L^\infty(\Omega)$, which fulfills $\underline{\kappa}\le\kappa(x)\le\overline{\kappa}$ for almost every $x\in\Omega$ with $0<\underline{\kappa}\le\overline{\kappa}<\infty$. The considered semilinear, elliptic PDE is

$$-\operatorname{div}(\kappa\nabla y)+\varphi(y)=Du+f \quad (\text{in }\Omega), \qquad y=0 \quad (\text{on }\partial\Omega), \tag{2.4}$$

where $\varphi:\mathbb{R}\to\mathbb{R}$ is twice continuously differentiable and monotonically increasing. Its second derivative shall fulfill the growth condition

$$|\varphi''(t)|\le a''_\varphi+c''_\varphi|t|^{p-3} \tag{2.5}$$

for all $t\in\mathbb{R}$ with constants $a''_\varphi,c''_\varphi\ge 0$ and the exponent $p\in(3,\infty)$ if $n=2$ and $p\in(3,6]$ if $n=3$. With $Y^*=H^{-1}(\Omega)$ and the definitions

$$\begin{aligned} &\langle Ay,\,v\rangle_{Y^*,Y}=(\kappa\nabla y,\nabla v)_{L^2(\Omega)^n}, \quad \langle N(y),\,v\rangle_{Y^*,Y}=\int_\Omega\varphi(y)v\,\mathrm{d}x,\\ &\langle Bu,\,v\rangle_{Y^*,Y}=(Du,v)_{L^2(\Omega)}, \qquad \langle b,v\rangle_{Y^*,Y}=(f,v)_{L^2(\Omega)} \end{aligned} \tag{2.6}$$

for every $v\in H_0^1(\Omega)$, the weak formulation of (2.4) is given by (2.1).

Under the standing assumptions, this state equation has a unique weak solution $y=S(u)\in Y\cap\mathscr{C}(\overline{\Omega})$, see, e.g., [14, Thm. 1.25, Remark 1.12]. The control-to-state mapping $S:U\to Y$ is Lipschitz continuous with constant $\frac{C_\Omega\|D\|_{\mathscr{L}(U,L^2(\Omega))}}{\underline{\kappa}}$, where C_Ω comes from the embedding $L^2(\Omega)\hookrightarrow H^{-1}(\Omega)$, and twice continuously differentiable. Furthermore, the optimal control problem (2.3) admits a solution, cf. [19, Lem. 9.4].

2.2. Derivatives of the reduced objective function. Since the state equation has a unique solution, (2.3) can be reduced to the control giving

$$\min_{u\in U}\hat J(u):=J(S(u),u) \quad \text{s.t.} \quad u\in U_{\mathrm{ad}}. \tag{2.7}$$

We use the adjoint approach [14, sec. 1.6] for the computation of the gradient and the Hessian of $\hat J$. This works because the operator $A+N'(y)$ is boundedly invertible for every $y\in Y$. The adjoint equation

$$Az+N'(y)z=-Q^*(Qy-\hat q) \tag{2.8}$$

with $y=S(u)$ has a unique solution $z=T(u)$. The derivatives of the reduced objective function $\hat J$ are given by

$$\langle\hat J'(u),w\rangle_{U^*,U}=\langle T(u),-Bw\rangle_{Y^*,Y}+\gamma(u,w)_U \quad \forall\, w\in U, \tag{2.9}$$

$$\nabla^2\hat J(u)s=\imath B^*h+\gamma s\in U, \tag{2.10}$$

where $h \in Y$ solves (2.11) with $\delta = S'(u)s \in Y$ solving (2.12):

$$[A + N'(S(u))]^* h = Q^* Q\delta + \langle T(u), [N''(S(u))\delta]\,\cdot\rangle_{Y,Y^*} \tag{2.11}$$

$$[A + N'(S(u))]\delta = Bs. \tag{2.12}$$

Here, $\iota : U^* \to U$ is the Riesz map and it is used that N is twice continuously differentiable. Both δ and h are uniquely determined.

3. An inexact trust-region algorithm. To solve (2.7), we present an inexact and projection-based trust-region algorithm, which can be used for the adaptive solution of optimization problems of the form

$$\min_{u\in U} \hat{J}(u) \quad \text{s.t.} \quad u \in U_{\text{ad}}, \tag{3.1}$$

and prove its convergence. It generalizes a version of the algorithms presented in [15, 16] and extends them to constrained problems using a possibly inexactly computed projection onto the feasible set $U_{\text{ad}} \subset U$. This is inspired by [22].

Assumption 3.1. We make the following assumptions on problem (3.1):

- U is a Hilbert space.
- The feasible set $U_{\text{ad}} \subset U$ is nonempty, closed, and convex.
- The objective function $\hat{J} : \tilde{U} \to \mathbb{R}$ is continuously differentiable on an open set $\tilde{U}$, $U_{\text{ad}} \subset \tilde{U} \subset U$, and bounded from below on U_{ad}. The Fréchet approximation condition holds uniformly on every level set, i.e., for every $\tilde{u} \in U_{\text{ad}}$ we have

$$\sup_{u\in U_{\text{ad}}:\hat{J}(u)\le\hat{J}(\tilde{u})} |\hat{J}(u+s) - \hat{J}(u) - (\nabla\hat{J}(u), s)_U| = o(\|s\|_U) \quad (s \to 0). \tag{3.2}$$

3.1. Formulation of the algorithm. For a comprehensive introduction to trust-region algorithms we refer to [5]. In each iteration k of the algorithm presented here, we use a typically, but not necessarily quadratic model $m_k(s)$ of $\hat{J}(u^k + s) - \hat{J}(u^k)$ with the current control u^k for the computation of the step $s^k \in U$. The step computation approximately solves

$$\min_{s\in U} m_k(s) \quad \text{s.t.} \quad u^k + s \in U_{\text{ad}}, \quad \|s\|_U \le \Delta_k \tag{3.3}$$

with the current trust region radius $\Delta_k > 0$. For the acceptance of the step, we allow for inexact evaluations of $\hat{J}$ by using an approximation $\hat{J}_k$ instead of $\hat{J}$. We define the actual, computed, and predicted reduction, respectively, as

$$\text{ared}_k := \hat{J}(u^k) - \hat{J}(u^k + s^k),\ \text{cred}_k := \hat{J}_k(u^k) - \hat{J}_k(u^k + s^k),\ \text{pred}_k := m_k(0) - m_k(s^k). \tag{3.4}$$

Furthermore, we define a criticality measure for the original problem (3.1), namely

$$\chi : \tilde{U} \to \mathbb{R}_{\ge 0}, \quad \chi(u) := \|u - P_{U_{\text{ad}}}(u - \tau\nabla\hat{J}(u))\|_U \tag{3.5}$$

with a fixed parameter $\tau > 0$ and the projection $P_{U_{\text{ad}}}$ onto the feasible set. The function χ is continuous, and $\chi(\bar{u}) = 0$ holds if and only if $\bar{u}$ is a first order critical point for (3.1). In addition, a criticality measure for problem (3.3) without the trust-region constraint is defined:

$$\tilde{\chi}_k : U \to \mathbb{R}_{\ge 0}, \quad \tilde{\chi}_k(s) := \|u^k + s - P_{U_{\text{ad}}}(u^k + s - \tau\nabla m_k(s))\|_U. \tag{3.6}$$

The condition $\tilde{\chi}_k(\bar{s}) = 0$ holds if and only if $\bar{s}$ is first order critical for the problem

$$\min_{s\in U} m_k(s) \quad \text{s.t.} \quad u^k + s \in U_{\text{ad}}. \tag{3.7}$$

Typically, the projection is also not computed exactly. Therefore, we introduce the *approximate criticality measure* for problem (3.7),

$$\chi_k : U \to \mathbb{R}_{\geq 0}, \quad \chi_k(s) := \|u^k + s - \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(u^k + s - \tau \nabla m_k(s))\|_U, \tag{3.8}$$

now with an *approximate* projection $\hat{\mathrm{P}}_{U_{\mathrm{ad}}} : U \to U_{\mathrm{ad}}$ onto the feasible set. This can be *any* mapping approximating the exact projection. Especially, it does not have to fulfill the variational inequality defining the projection on some discrete subspace $\mathrm{U} \subset U$. In contrast to that, the properties of the discrete projection are used in [22, Lem. 5.3, Lem. 5.5] when proving the Cauchy decrease condition although an approximate version is used in the final implementation there [22, sec. 5.2].

To ensure global convergence of the algorithm we have to control the following quantities:

- The inexactness of the model gradient:

$$\|\nabla m_k(0) - \nabla \hat{J}(u^k)\|_U \leq \rho_{\mathrm{g}}(\Delta_k), \tag{3.9}$$

  where $\rho_{\mathrm{g}} : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ is a function satisfying $\lim_{t \to 0^+} \rho_{\mathrm{g}}(t) = 0$, e.g., $\rho_{\mathrm{g}}(t) = \mathfrak{c}_{\mathrm{g}} t$, $\mathfrak{c}_{\mathrm{g}} > 0$.
- The inexactness of the approximate criticality measure:

$$|\chi_k(0) - \chi(u^k)| \leq \rho_{\mathrm{c}}(\chi_k(0)), \tag{3.10}$$

  where $\rho_{\mathrm{c}} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a function satisfying $\lim_{t \to 0^+} \rho_{\mathrm{c}}(t) = 0$ and $\rho_{\mathrm{c}}(0) = 0$, e.g., $\rho_{\mathrm{c}}(t) = \mathfrak{c}_{\mathrm{c}} t$, $\mathfrak{c}_{\mathrm{c}} > 0$.
- The quality of the computed reduction:

$$|\mathrm{ared}_k - \mathrm{cred}_k| \leq \rho_{\mathrm{r}}(\eta_3 \min\{\mathrm{pred}_k, \mathfrak{r}_k\}), \tag{3.11}$$

  with $\eta_3 < \min\{\eta_1, 1 - \eta_2\}$, where $0 < \eta_1 < \eta_2 < 1$ are a priori chosen parameters for assessing the quality of the model, a forcing sequence $(\mathfrak{r}_k)_{k \in \mathbb{N}_0} \subset \mathbb{R}_{>0}$ fulfilling $\lim_{k \to \infty} \mathfrak{r}_k = 0$, and a function $\rho_{\mathrm{r}} : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ fulfilling $\rho_{\mathrm{r}}(t) \leq t$ for all $t \in (0, \bar{t}]$ with some fixed $\bar{t} > 0$, e.g., $\rho_{\mathrm{r}}(t) = \mathfrak{c}_{\mathrm{r}} t^{\mathfrak{e}_{\mathrm{r}}}$, $\mathfrak{c}_{\mathrm{r}} > 0$, $\mathfrak{e}_{\mathrm{r}} > 1$. Note that $\rho_{\mathrm{r}}(t) = t$ would also be possible, but then it is not sufficient to know the error in (3.11) up to an unknown, multiplicative constant.

A trial step $s^k \in U_{\mathrm{ad}} - u^k$, $\|s^k\|_U \leq \Delta_k$, has to fulfill the decrease condition

$$\mathrm{pred}_k = m_k(0) - m_k(s^k) \geq \rho_{\mathrm{t1}}(\chi_k(0)) \cdot \min\{\rho_{\mathrm{t2}}(\chi_k(0)), \Delta_k\} \tag{3.12}$$

with monotonically increasing functions $\rho_{\mathrm{t1}}, \rho_{\mathrm{t2}} : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$. These functions must be chosen such that (3.12) is satisfiable by, e.g., a generalized Cauchy point, see subsection 3.3. A possible example is $\rho_{\mathrm{t1}}(t) = \mathfrak{c}_{\mathrm{t1}} t$, $\rho_{\mathrm{t1}}(t) = \mathfrak{c}_{\mathrm{t2}} t$ with $\mathfrak{c}_{\mathrm{t1}}, \mathfrak{c}_{\mathrm{t2}} > 0$.

The complete method is listed in Algorithm 1. All iterates u^k belong to U_{ad} since $s^k \in U_{\mathrm{ad}} - u^k$ is required for all trial steps. Therefore, it is sufficient to assume differentiability of $\hat{J}$ only in an open neighborhood $\tilde{U}$ of U_{ad}, see Assumption 3.1.

3.2. Convergence proof. Provided all conditions in Algorithm 1 can be satisfied, we prove its convergence. This means that we assume for now that an adequate model, approximate projection, trial step, and inexact objective function exist in each iteration. We discuss this in subsection 3.3.

We require the following assumption in addition to Assumption 3.1 to prove the convergence result given in Theorem 3.3.

Assumption 3.2. Each model $m_k : U \to \mathbb{R}$ is continuously differentiable and fulfills

$$\sup_{k \in \mathbb{N}_0} |m_k(s) - m_k(0) - (\nabla m_k(0), s)_U| = o(\|s\|) \quad (s \to 0), \tag{3.13}$$

$$\|\nabla m_k(s) - \nabla m_k(\hat{s})\|_U \leq \mathfrak{c}_{m_k} \|s - \hat{s}\|_U \tag{3.14}$$

Algorithm 1: Inexact trust-region method for solving problem (3.1)

Input: Initial iterate $u^0 \in U_{\mathrm{ad}}$

Parameters : $\tau > 0$, error control functions ρ_{c}, ρ_{g}, ρ_{t1}, ρ_{t2}, ρ_{r}, forcing sequence $(\mathfrak{r}_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}_{>0}$, $\lim_{k\to\infty} \mathfrak{r}_k = 0$, $\Delta_{\max} \in (0,\infty]$, $\Delta_0 \in \mathbb{R}_{>0}$ s.t. $\Delta_0 \le \Delta_{\max}$, $0 < \eta_1 < \eta_2 < 1$ and $0 < \eta_3 \le \min\{\eta_1, 1-\eta_2\}$, $0 < \nu_1 < 1 \le \nu_2 < \nu_3$.

Output: Sequences $(u^k)_{k\in\mathbb{N}_0} \subset U_{\mathrm{ad}}$, $(\Delta_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}_{>0}$, $(\chi_k(0))_{k\in\mathbb{N}_0} \subset \mathbb{R}_{\ge 0}$

for $k := 0, 1, 2, \ldots$ **do**

Choose a model $m_k : U \to \mathbb{R}$ and an approximate projection $\hat{\mathrm{P}}_{U_{\mathrm{ad}}} : U \to U_{\mathrm{ad}}$ such that (3.9) and (3.10) hold. Compute $\chi_k(0)$ using m_k and $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$.

if $\chi_k(0) = 0$, **then**

set $u^\ell := u^k$, $\Delta_\ell := \Delta_k$, and $\chi_\ell(0) = 0$ for all $\ell \ge k+1$ and STOP.

Compute a trial step $s^k \in U_{\mathrm{ad}} - u^k$, $\|s^k\|_U \le \Delta_k$, fulfilling (3.12) with the computed $\chi_k(0)$, see subsection 3.3.

Compute pred_k and cred_k by (3.4) with $\hat{J}_k$ such that (3.11) holds.

if $\frac{\mathrm{cred}_k}{\mathrm{pred}_k} < \eta_1$ *(unsuccessful step),* **then**

$u^{k+1} := u^k$, choose $\Delta_{k+1} \in (0, \nu_1\Delta_k]$.

else if $\frac{\mathrm{cred}_k}{\mathrm{pred}_k} \in [\eta_1, \eta_2)$ *(successful step),* **then**

$u^{k+1} := u^k + s^k$, choose $\Delta_{k+1} \in [\nu_1\Delta_k, \min\{\nu_2\Delta_k, \Delta_{\max}\}]$.

else if $\frac{\mathrm{cred}_k}{\mathrm{pred}_k} \ge \eta_2$ *(very successful step),* **then**

$u^{k+1} := u^k + s^k$, choose $\Delta_{k+1} \in [\min\{\nu_2\Delta_k, \Delta_{\max}\}, \min\{\nu_3\Delta_k, \Delta_{\max}\}]$.

for all $s, \hat{s} \in U_{\mathrm{ad}} - u^k$, $\|s\|_U \le \Delta_k$, $\|\hat{s}\|_U \le \Delta_k$ with some $\mathfrak{c}_{m_k} > 0$. This means that the Fréchet approximation condition holds uniformly over all models and the model gradients are Lipschitz continuous on the feasible search directions. The Lipschitz constants shall be bounded uniformly: $\mathfrak{c}_{m_k} \le \mathfrak{c}_m$ for some $\mathfrak{c}_m > 0$ and all $k \in \mathbb{N}_0$.

THEOREM 3.3. *Let Assumptions 3.1 and 3.2 hold and let the sequence $(u^k)_{k\in\mathbb{N}_0} \subset U$ be generated by Algorithm 1. Then,* $\liminf_{k\to\infty} \chi(u^k) = 0$ *holds with the criticality measure χ defined in (3.5).*

Proof. The proof from [16] can be adapted, see Appendix A. ∎

3.3. Satisfying the conditions required by the algorithm. We show that Algorithm 1 is realizable, i.e., that all requirements can be met under certain assumptions. This includes the computation of a generalized Cauchy point satisfying (3.12). We use the proposed error functions $\rho_{\mathrm{g}}(t) = \mathfrak{c}_{\mathrm{g}} t$ etc. and suppose that we can compute the objective function, its gradient, and the projection onto the feasible set to any given accuracy in each iteration k. Then, we have bounds of the form

$$\begin{aligned} &\text{(a)}\ |\hat{J}_k(u) - \hat{J}(u)| \le c_{\mathrm{o}}\varepsilon_{\mathrm{o}}, \qquad \text{(b)}\ \|\nabla m_k(0) - \nabla\hat{J}(u^k)\|_U \le c_{\mathrm{g}}\varepsilon_{\mathrm{g}}, \\ &\text{(c)}\ \|P_{U_{\mathrm{ad}}}(w^k(t)) - \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t))\|_U \le c_{\mathrm{p}}\varepsilon_{\mathrm{p}} \end{aligned} \tag{3.15}$$

with arbitrarily small, computable tolerances $\varepsilon_{\mathrm{o}}, \varepsilon_{\mathrm{g}}, \varepsilon_{\mathrm{p}} > 0$. The constants $c_{\mathrm{o}}, c_{\mathrm{g}}, c_{\mathrm{p}} > 0$ are fixed, but possibly unknown, as it can happen in the PDE context when applying a posteriori error estimates. Here, $u^k \in U_{\mathrm{ad}}$ is the current iterate, and $u = u^k$ or $u = u^k + s^k$ with s^k being the trial step. Furthermore, $w^k(t) := u^k - t\nabla m_k(0) \in U$ with a suitable gradient stepsize $t > 0$.

Model and approximate projection. We ensure (3.9) and (3.10) by computing the gradient and the projection accurately enough, i.e., we choose arbitrary constants $\tilde{\mathfrak{c}}_{\mathrm{c}}, \tilde{\mathfrak{c}}_{\mathrm{g}}, \tilde{\mathfrak{c}}_{\mathrm{p}} > 0$

and require (3.15) (b,c) with the computable bounds

$$\text{(a)}\ \ \varepsilon_{\mathrm{g}} \le \min\{\tilde{\mathfrak{c}}_{\mathrm{c}}\chi_k(0), \tilde{\mathfrak{c}}_{\mathrm{g}}\Delta_k\}, \quad \text{(b)}\ \ \varepsilon_{\mathrm{p}} \le \tilde{\mathfrak{c}}_{\mathrm{p}}\chi_k(0). \tag{3.16}$$

PROPOSITION 3.4. *Let (3.16) and (3.15) (b,c) with $t = \tau$ hold true. Then, (3.9) and (3.10) follow with $\rho_{\mathrm{g}}(t) = \mathfrak{c}_{\mathrm{g}}t$, $\mathfrak{c}_{\mathrm{g}} = c_{\mathrm{g}}\tilde{\mathfrak{c}}_{\mathrm{g}}$, and $\rho_{\mathrm{c}}(t) = \mathfrak{c}_{\mathrm{c}}t$, $\mathfrak{c}_{\mathrm{c}} = c_{\mathrm{p}}\tilde{\mathfrak{c}}_{\mathrm{p}} + \tau c_{\mathrm{g}}\tilde{\mathfrak{c}}_{\mathrm{c}}$, respectively.*

Proof. The estimate (3.9) follows directly by combining (3.15) (b) and (3.16) (a). Using the definitions (3.5), (3.6), and (3.8), we estimate

$$\begin{aligned}
|\chi_k(0) - \chi(u^k)| &\le |\chi_k(0) - \tilde{\chi}_k(0)| + |\tilde{\chi}_k(0) - \chi(u^k)| \\
&\le \|P_{U_{\mathrm{ad}}}(u^k - \tau\nabla m_k(0)) - \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(u^k - \tau\nabla m_k(0))\|_U \\
&\quad + \|P_{U_{\mathrm{ad}}}(u^k - \tau\nabla\hat{J}(u^k)) - P_{U_{\mathrm{ad}}}(u^k - \tau\nabla m_k(0))\|_U \\
&\le \|[P_{U_{\mathrm{ad}}} - \hat{\mathrm{P}}_{U_{\mathrm{ad}}}](u^k - \tau\nabla m_k(0))\|_U + \tau\|\nabla m_k(0) - \nabla\hat{J}(u^k)\|_U \\
&\le c_{\mathrm{p}}\tilde{\mathfrak{c}}_{\mathrm{p}}\chi_k(0) + \tau c_{\mathrm{g}}\tilde{\mathfrak{c}}_{\mathrm{c}}\chi_k(0) = \rho_{\mathrm{c}}(\chi_k(0)),
\end{aligned} \tag{3.17}$$

where the second inequality is established using $\big|\|u\|_U - \|w\|_U\big| \le \|u - w\|_U$ for any $u, w \in U$ and the last one follows from (3.15) (b,c) and (3.16). This proves that condition (3.10) is satisfied. ∎

If $\chi(u^k) > 0$, condition (3.16) is fulfilled if $\varepsilon_{\mathrm{g}}, \varepsilon_{\mathrm{p}} > 0$ are small enough: Then, the left-hand sides of the inequalities are close enough to zero and the right-hand sides are close to $\min\{\tilde{\mathfrak{c}}_{\mathrm{c}}\chi(u^k), \tilde{\mathfrak{c}}_{\mathrm{g}}\Delta_k\} > 0$ and $\tilde{\mathfrak{c}}_{\mathrm{p}}\chi(u^k) > 0$, respectively, because $\chi_k(0) \to \chi(u^k)$ as $(\varepsilon_{\mathrm{g}}, \varepsilon_{\mathrm{p}}) \to 0$ by (3.15) (b,c). In the case $\chi(u^k) = 0$ it could happen that an adaptive algorithm for computing $\nabla m_k(0)$ and the projection keeps increasing the accuracy, i.e., decreasing ε_{g} and ε_{p} towards 0, without being able to fulfill (3.16). For theoretical considerations, we assume that the exact gradient and projection are used in this case. In a practical implementation, the algorithm should be stopped if ε_{g}, ε_{p}, and $\chi_k(0)$ are sufficiently small.

Generalized Cauchy point. Condition (3.12) will be satisfied by a generalized Cauchy point $s_{\mathrm{C}}^k \in U_{\mathrm{ad}} - u^k$, $\|s_{\mathrm{C}}^k\|_U \le \Delta_k$, which is computed by a projected linesearch with an inexact, and possibly refined projection $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$, cf. [22]. It is very important to permit an inexact projection in this procedure because then U-grid refinement may not be necessary in every iteration. In this way, computational cost can be saved and the quality of the FE grid can be preserved by a suitable refinement method instead of adapting the refinement exactly to the projection which has to be computed.

In the following, we assume that the inexact criticality measure is computed using the exact projection, i.e., $\tilde{\chi}_k(0) = \chi_k(0)$. Since $\chi_k(0) > 0$ is ensured by the stopping criterion of Algorithm 1, this means that $s = 0$ is not stationary for (3.7) when the linesearch is performed. Moreover, we can choose $\tilde{\mathfrak{c}}_{\mathrm{p}} = 0$ in (3.16) (b). In our final implementation, this is not a drawback because the exact projection has to be computed nevertheless to evaluate the projection error.

LEMMA 3.5. *Given $u^k \in U_{\mathrm{ad}}$, $\nabla m_k(0) \in U$, and $t > 0$, the direction*

$$p^k(t) := \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t)) - u^k \quad \text{with } w^k(t) = u^k - t\nabla m_k(0) \tag{3.18}$$

is a descent direction for m_k in 0 in the sense that $(\nabla m_k(0), p^k(t))_U \le -\frac{\mathfrak{c}_{\mathrm{i}}}{t}\|p^k(t)\|_U^2$, provided the inexact projection $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$ satisfies

$$\big(\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t)) - w^k(t), \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t)) - u^k\big)_U \le (1 - \mathfrak{c}_{\mathrm{i}})\|p^k(t)\|_U^2 \tag{3.19}$$

for some arbitrary, but fixed constant $\mathfrak{c}_{\mathrm{i}} \in (0, 1]$ and $\|p^k(t)\|_U > 0$.

Proof. Writing $w^k = w^k(t)$ and $p^k = p^k(t)$, we estimate

$$\begin{aligned} t\big(\nabla m_k(0), p^k\big)_U &= \big(u^k - w^k, \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k) - u^k\big)_U \\ &= -\|p^k\|_U^2 + \big(\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k) - w^k, \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k) - u^k\big)_U \leq -\mathfrak{c}_{\mathrm{i}}\|p^k\|_U^2. \end{aligned} \tag{3.20}$$

Due to $t > 0$, $\mathfrak{c}_{\mathrm{i}} > 0$ and $\|p^k\| > 0$, p^k is a descent direction for m_k in 0. ∎

Lemma 3.5 provides the computable condition (3.19) on the approximate projection $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$ to make the approximately projected negative gradient a descent direction. It is essentially different from [22, Lem. 5.3] in the sense that we do not use any projection property of $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$. If the discrete projection is used, (3.19) is trivially fulfilled. If not, it is sufficient to compute the projection accurately enough, i.e., to take $\varepsilon_{\mathrm{p}} > 0$ small enough in (3.15) (c): Since $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t)) \to P_{U_{\mathrm{ad}}}(w^k(t))$ as $\varepsilon_{\mathrm{p}} \to 0$, the left-hand side in (3.19) is then close enough to a non-positive real number. Furthermore, $\|p^k(t)\|_U \to \|P_{U_{\mathrm{ad}}}(w^k(t)) - u^k\|_U > 0$ as $\varepsilon_{\mathrm{p}} \to 0$. The positiveness for every $t > 0$ follows from $\tilde{\chi}_k(0) > 0$: If there existed $t > 0$ such that $\|P_{U_{\mathrm{ad}}}(w^k(t)) - u^k\|_U = 0$, then $0 \in U$ would be first-order stationary for problem (3.7), which would yield $\tilde{\chi}_k(0) = 0$.

Definition 3.6 (Armijo condition). Choose the largest $t_k \in \{\mathfrak{c}_{\mathrm{a},k} \cdot \mathfrak{c}_{\mathrm{f}}^j, j \in \mathbb{N}_0\}$ such that

$$\text{(a) } m_k(p^k(t_k)) \leq m_k(0) - \tfrac{\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{e}}}{t_k}\|p^k(t_k)\|_U^2, \qquad \text{(b) } \|p^k(t_k)\|_U \leq \Delta_k \tag{3.21}$$

with two fixed parameters $\mathfrak{c}_{\mathrm{f}}, \mathfrak{c}_{\mathrm{e}} \in (0,1)$ are satisfied. Here, $\mathfrak{c}_{\mathrm{a},k} \in \mathbb{R}$ are constants satisfying $\mathfrak{c}_{\mathrm{a}} \leq \mathfrak{c}_{\mathrm{a},k} \leq \tau$ for all $k \in \mathbb{N}_0$ with some $\mathfrak{c}_{\mathrm{a}} \in (0, \tau]$.

Lemma 3.7. *If Assumption 3.2 is true and if $p^k(t_k)$ is computed according to Lemma 3.5, condition (3.21) (a) is satisfied for all $t_k \in \big(0, \frac{2(1-\mathfrak{c}_{\mathrm{e}})\mathfrak{c}_{\mathrm{i}}}{\mathfrak{c}_{m_k}}\big]$, where $\mathfrak{c}_{m_k} > 0$ is the Lipschitz constant of the model gradient ∇m_k.*

Proof. We estimate using the fundamental theorem of calculus:

$$\begin{aligned} &m_k(p^k(t_k)) - m_k(0) = \int_0^1 \big(\nabla m_k(\sigma p^k(t_k)), p^k(t_k)\big)_U \,\mathrm{d}\sigma \\ &\leq \big(\nabla m_k(0), p^k(t_k)\big)_U + \int_0^1 \big\|\nabla m_k(\sigma p^k(t_k)) - \nabla m_k(0)\big\|_U \cdot \big\|p^k(t_k)\big\|_U \,\mathrm{d}\sigma \\ &\leq -\tfrac{\mathfrak{c}_{\mathrm{i}}}{t_k}\|p^k(t_k)\|_U^2 + \tfrac{\mathfrak{c}_{m_k}}{2}\|p^k(t_k)\|_U^2 = \big(-\tfrac{\mathfrak{c}_{\mathrm{i}}}{t_k} + \tfrac{\mathfrak{c}_{m_k}}{2}\big)\|p^k(t_k)\|_U^2. \end{aligned}$$

In the last estimate we have used Lemma 3.5 and the Lipschitz continuity of ∇m_k. With the given choice of t_k, (3.21) (a) follows. ∎

To show (3.12), the inexact projection has to be accurate enough during linesearch, i.e., we choose two constants $\tilde{\mathfrak{c}}_{\mathrm{l}1}, \tilde{\mathfrak{c}}_{\mathrm{l}2} \in (0, \infty)$ and require (3.15) (c) with

$$\text{(a) } \varepsilon_{\mathrm{p}} \leq \tilde{\mathfrak{c}}_{\mathrm{l}1}\|\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t)) - u^k\|_U, \qquad \text{(b) } \varepsilon_{\mathrm{p}} \leq \tilde{\mathfrak{c}}_{\mathrm{l}2}\|P_{U_{\mathrm{ad}}}(w^k(t)) - u^k\|_U. \tag{3.22}$$

Lemma 3.8. *Let Assumption 3.2 hold and let t_k be computed according to the Armijo condition (Definition 3.6), where the inexact projection $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$ satisfies (3.19) and additionally (3.15) (c) with (3.22) for every $t = t_k$ tested during the Armijo linesearch. Define $\mathfrak{c}_{\mathrm{l}1} := \frac{1}{c_{\mathrm{p}}\tilde{\mathfrak{c}}_{\mathrm{l}1}+1} \in (0,1]$ and $\mathfrak{c}_{\mathrm{l}2} := \frac{1}{c_{\mathrm{p}}\tilde{\mathfrak{c}}_{\mathrm{l}2}+1} \in (0,1]$. Then, the trial step $s_{\mathrm{C}}^k := p^k(t_k)$ fulfills condition (3.12) with $\rho_{\mathrm{t}1}(t) = \mathfrak{c}_{\mathrm{t}1}t$, $\mathfrak{c}_{\mathrm{t}1} = \frac{\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{l}1}^2\mathfrak{c}_{\mathrm{l}2}\mathfrak{c}_{\mathrm{e}}\mathfrak{c}_{\mathrm{f}}}{\tau}$, $\rho_{\mathrm{t}1}(t) = \mathfrak{c}_{\mathrm{t}2}t$, $\mathfrak{c}_{\mathrm{t}2} = \frac{1}{\mathfrak{c}_{\mathrm{l}2}\tau}\min\big\{\frac{2(1-\mathfrak{c}_{\mathrm{e}})\mathfrak{c}_{\mathrm{i}}}{\mathfrak{c}_m}, \frac{\mathfrak{c}_{\mathrm{a}}}{\mathfrak{c}_{\mathrm{f}}}\big\}$, and $\chi_k(0) = \tilde{\chi}_k(0)$.*

Proof. From (3.15) (c), (3.22) (a) we get that

$$\begin{aligned}\mathfrak{c}_{l1}\|P_{U_{\mathrm{ad}}}(w^k(t))-u^k\|_U &\le \mathfrak{c}_{l1}\|P_{U_{\mathrm{ad}}}(w^k(t))-\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t))\|_U+\mathfrak{c}_{l1}\|\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t))-u^k\|_U\\ &\le \mathfrak{c}_{l1}(c_{\mathrm{p}}\tilde{\mathfrak{c}}_{l1}+1)\|\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t))-u^k\|_U=\|p^k(t)\|_U\end{aligned}\tag{3.23}$$

for every tested $t=t_k$. Using

$$\tfrac{1}{t}\|P_{U_{\mathrm{ad}}}(w^k(t))-u^k\|_U\ge\tfrac{1}{\tau}\tilde{\chi}_k(0)\tag{3.24}$$

for $t\le\tau$ by [14, Lem. 1.10 (e)], (3.21) (a), and (3.23), we conclude that

$$\begin{aligned}\mathrm{pred}_k=m_k(0)-m_k(p^k(t_k))&\ge\tfrac{\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{e}}}{t_k}\|p^k(t_k)\|_U^2\\ &\ge\tfrac{\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{e}}\mathfrak{c}_{l1}}{t_k}\|P_{U_{\mathrm{ad}}}(w^k(t_k))-u^k\|_U\,\|p^k(t_k)\|_U\ge\tfrac{\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{e}}\mathfrak{c}_{l1}}{\tau}\tilde{\chi}_k(0)\,\|p^k(t_k)\|_U\end{aligned}\tag{3.25}$$

holds by $t_k\le\mathfrak{c}_{\mathrm{a},k}\le\tau$.
Now consider the case that t_k found by the standard projected Armijo linesearch for (3.21) (a) already satisfies (3.21) (b). Thus, Lemma 3.7 can be applied and $t_k\ge\min\{\frac{2(1-\mathfrak{c}_{\mathrm{e}})\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{f}}}{\mathfrak{c}_{m_k}},\mathfrak{c}_{\mathrm{a},k}\}$ holds. From (3.25), (3.23), and (3.24) it now follows that

$$\begin{aligned}\mathrm{pred}_k&\ge\tfrac{\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{e}}\mathfrak{c}_{l1}}{\tau}\tilde{\chi}_k(0)\,\tfrac{\mathfrak{c}_{l1}t_k}{\tau}\tilde{\chi}_k(0)\ge\tfrac{\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{e}}\mathfrak{c}_{l1}}{\tau}\tilde{\chi}_k(0)\min\{\tfrac{2(1-\mathfrak{c}_{\mathrm{e}})\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{f}}}{\mathfrak{c}_{m_k}},\mathfrak{c}_{\mathrm{a},k}\}\,\tfrac{\mathfrak{c}_{l1}}{\tau}\tilde{\chi}_k(0)\\ &=\tfrac{\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{l1}^2\mathfrak{c}_{l2}\mathfrak{c}_{\mathrm{e}}\mathfrak{c}_{\mathrm{f}}}{\tau}\tilde{\chi}_k(0)\,\tfrac{1}{\mathfrak{c}_{l2}\tau}\min\{\tfrac{2(1-\mathfrak{c}_{\mathrm{e}})\mathfrak{c}_{\mathrm{i}}}{\mathfrak{c}_m},\tfrac{\mathfrak{c}_{\mathrm{a}}}{\mathfrak{c}_{\mathrm{f}}}\}\tilde{\chi}_k(0)\\ &=\rho_{\mathrm{t}1}(\tilde{\chi}_k(0))\,\rho_{\mathrm{t}2}(\tilde{\chi}_k(0))\ge\rho_{\mathrm{t}1}(\tilde{\chi}_k(0))\min\{\rho_{\mathrm{t}2}(\tilde{\chi}_k(0)),\Delta_k\}.\end{aligned}$$

In the case that the standard search for (3.21) (a) does not yield t_k satisfying (3.21) (b), t_k has to be decreased further. It follows that $\big\|p^k(\frac{t_k}{\mathfrak{c}_{\mathrm{f}}})\big\|_U>\Delta_k$.
In analogy to (3.23) we can conclude from (3.15) (c), (3.22) (b) that

$$\mathfrak{c}_{l2}\|p^k(t)\|_U=\mathfrak{c}_{l2}\|\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t))-u^k\|_U\le\|P_{U_{\mathrm{ad}}}(w^k(t))-u^k\|_U\tag{3.26}$$

for every tested $t=t_k$. With (3.23), [14, Lem. 1.10 (e)], and (3.26) we obtain

$$\begin{aligned}\|p^k(t_k)\|_U&\ge\mathfrak{c}_{l1}\|P_{U_{\mathrm{ad}}}(w^k(t_k))-u^k\|_U\ge\mathfrak{c}_{l1}\mathfrak{c}_{\mathrm{f}}\|P_{U_{\mathrm{ad}}}(w^k(\tfrac{t_k}{\mathfrak{c}_{\mathrm{f}}}))-u^k\|_U\\ &\ge\mathfrak{c}_{l1}\mathfrak{c}_{l2}\mathfrak{c}_{\mathrm{f}}\|p^k(\tfrac{t_k}{\mathfrak{c}_{\mathrm{f}}})\|_U>\mathfrak{c}_{l1}\mathfrak{c}_{l2}\mathfrak{c}_{\mathrm{f}}\Delta_k.\end{aligned}\tag{3.27}$$

Hence, by (3.25), we get

$$\begin{aligned}\mathrm{pred}_k&\ge\tfrac{\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{e}}\mathfrak{c}_{l1}}{\tau}\tilde{\chi}_k(0)\,\|p^k(t_k)\|_U>\tfrac{\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{e}}\mathfrak{c}_{l1}}{\tau}\tilde{\chi}_k(0)\,\mathfrak{c}_{l1}\mathfrak{c}_{l2}\mathfrak{c}_{\mathrm{f}}\Delta_k\\ &=\rho_{\mathrm{t}1}(\tilde{\chi}_k(0))\,\Delta_k\ge\rho_{\mathrm{t}1}(\tilde{\chi}_k(0))\min\{\rho_{\mathrm{t}2}(\tilde{\chi}_k(0)),\Delta_k\}.\end{aligned}\tag{3.28}$$

Therefore, in both cases, (3.12) is satisfied with $\chi_k(0)=\tilde{\chi}_k(0)$. ∎

To guarantee (3.22) (b), we have to compute the exact projection or a suitable lower bound for $\|P_{U_{\mathrm{ad}}}(w^k(t))-u^k\|_U$. Again, condition (3.22) can be satisfied by taking $\varepsilon_{\mathrm{p}}>0$ small enough because the right-hand side is then sufficiently close to $\min\{\tilde{\mathfrak{c}}_{l1},\tilde{\mathfrak{c}}_{l2}\}\|P_{U_{\mathrm{ad}}}(w^k(t))-u^k\|_U>0$.

Remark 3.9.

- The proof of Lemma 3.8 shows that (3.22) (b) can be dropped during the linesearch for (3.21) (a).

- If (3.22) (a) holds with $\tilde{\mathfrak{c}}_{\mathrm{l}1} \in (0, \frac{1}{c_{\mathrm{p}}})$, (3.22) (b) follows with $\tilde{\mathfrak{c}}_{\mathrm{l}2} = \frac{1}{1-c_{\mathrm{p}}\tilde{\mathfrak{c}}_{\mathrm{l}1}} \in (0,1]$:

$$\begin{aligned}\|P_{U_{\mathrm{ad}}}(w^k(t)) - \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t))\|_U &\le c_{\mathrm{p}}\tilde{\mathfrak{c}}_{\mathrm{l}1}\|\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t)) - u^k\|_U \\ \le c_{\mathrm{p}}\tilde{\mathfrak{c}}_{\mathrm{l}1}\big(\|\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t)) - P_{U_{\mathrm{ad}}}(w^k(t))\|_U + \|P_{U_{\mathrm{ad}}}(w^k(t)) - u^k\|_U\big)\end{aligned}$$

  yields the result. For this purpose, the constant c_{p} has to be known. We follow this strategy in our implementation, where the projection error can be computed exactly.
- (3.22) (b) can be dropped if the discrete projection $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$ onto $U_{\mathrm{ad}} \cap \mathrm{U}$ is used (cf. [22]) since then $\|p^k(t_k)\|_U \ge \mathfrak{c}_{\mathrm{f}}\|p^k(\frac{t_k}{\mathfrak{c}_{\mathrm{f}}})\|_U$ holds by the projection property of $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$, which we do not assume in general, and then replaces (3.27). The constants in $\rho_{\mathrm{t}1}$ and $\rho_{\mathrm{t}2}$ change accordingly.
- Given $\mathfrak{c}_{\mathrm{a}} \in (0,\tau]$ and $\mathfrak{c}_{\mathrm{l}2}$, we choose $\mathfrak{c}_{\mathrm{a},k} := \max\{\mathfrak{c}_{\mathrm{a}}, \min\{\tau, \frac{\mathfrak{c}_{\mathrm{l}2}\Delta_k}{\|\nabla m_k(0)\|_U}\}\}$ in the Armijo condition because with $t_k \le \frac{\mathfrak{c}_{\mathrm{l}2}\Delta_k}{\|\nabla m_k(0)\|_U}$, which is not always ensured by the choice of $\mathfrak{c}_{\mathrm{a},k}$, and (3.26) we obtain

$$\begin{aligned}\|p^k(t_k)\|_U &\le \tfrac{1}{\mathfrak{c}_{\mathrm{l}2}}\|P_{U_{\mathrm{ad}}}(u^k - t_k\nabla m_k(0)) - u^k\|_U \\ &= \tfrac{1}{\mathfrak{c}_{\mathrm{l}2}}\|P_{U_{\mathrm{ad}}}(u^k - t_k\nabla m_k(0)) - P_{U_{\mathrm{ad}}}(u^k)\|_U \le \tfrac{t_k}{\mathfrak{c}_{\mathrm{l}2}}\|\nabla m_k(0)\|_U \le \Delta_k,\end{aligned}$$

  which is exactly (3.21) (b). Note that this choice of t_k is not necessary for (3.21) (b). This observation yields, together with Lemma 3.7, that the projected Armijo linesearch terminates after finitely many times decreasing t_k.

Based on the generalized Cauchy point s_{C}^k we get simple criterion for (3.12):

LEMMA 3.10 (Fraction of generalized Cauchy decrease). *Let s_{C}^k be computed according to Lemma 3.8 and let $s^k \in U_{\mathrm{ad}} - u^k$ satisfy*

$$m_k(0) - m_k(s^k) \ge \mathfrak{c}_{\mathrm{d}}(m_k(0) - m_k(s_{\mathrm{C}}^k)) \tag{3.29}$$

with some $\mathfrak{c}_{\mathrm{d}} \in (0,1]$. Then, s^k satisfies (3.12) with $\chi_k(0) = \tilde{\chi}_k(0)$ and

$$\rho_{\mathrm{t}1}(t) := \tfrac{\mathfrak{c}_{\mathrm{d}}\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{l}1}^2\mathfrak{c}_{\mathrm{l}2}\mathfrak{c}_{\mathrm{e}}\mathfrak{c}_{\mathrm{f}}}{\tau}t, \qquad \rho_{\mathrm{t}2}(t) := \tfrac{1}{\mathfrak{c}_{\mathrm{l}2}\tau}\cdot\min\big\{\tfrac{2(1-\mathfrak{c}_{\mathrm{e}})\mathfrak{c}_{\mathrm{i}}}{\mathfrak{c}_m}, \tfrac{\mathfrak{c}_{\mathrm{a}}}{\mathfrak{c}_{\mathrm{f}}}\big\}t.$$

Proof. From Lemma 3.8 we know that s_{C}^k satisfies (3.12) with

$$\rho_{\mathrm{t}1}(t) := \tfrac{\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{l}1}^2\mathfrak{c}_{\mathrm{l}2}\mathfrak{c}_{\mathrm{e}}\mathfrak{c}_{\mathrm{f}}}{\tau}t, \qquad \rho_{\mathrm{t}2}(t) := \tfrac{1}{\mathfrak{c}_{\mathrm{l}2}\tau}\cdot\min\big\{\tfrac{2(1-\mathfrak{c}_{\mathrm{e}})\mathfrak{c}_{\mathrm{i}}}{\mathfrak{c}_m}, \tfrac{\mathfrak{c}_{\mathrm{a}}}{\mathfrak{c}_{\mathrm{f}}}\big\}t,$$

and $\chi_k(0) = \tilde{\chi}_k(0)$. The stated result follows immediately. ∎

Computed reduction. The computed reduction is only evaluated as long as $\chi_k(0)$ is positive, which is ensured by the stopping criterion of Algorithm 1. Then the predicted reduction pred_k is positive by (3.12). The inequality (3.11) can be reduced to a bound on the inexact objective function evaluation:

LEMMA 3.11. *Let (3.15) (a) hold for all $u \in \{u^k, u^k + s^k\}$ with*

$$\varepsilon_{\mathrm{o}} \le \tilde{\mathfrak{c}}_{\mathrm{o}}(\eta_3 \min\{\mathrm{pred}_k, \mathfrak{r}_k\})^{\mathfrak{e}_{\mathrm{o}}} > 0,$$

where $\tilde{\mathfrak{c}}_{\mathrm{o}} > 0$, $\mathfrak{e}_{\mathrm{o}} > 1$ are chosen constants. Then, condition (3.11) holds with $\rho_{\mathrm{r}}(t) = 2c_{\mathrm{o}}\tilde{\mathfrak{c}}_{\mathrm{o}}t^{\mathfrak{e}_{\mathrm{o}}}$.

Proof. Using the definitions (3.4), (3.11) follows from

$$|\mathrm{ared}_k - \mathrm{cred}_k| \le |\hat{J}(u^k) - \hat{J}_k(u^k)| + |\hat{J}_k(u^k + s^k) - \hat{J}(u^k + s^k)|.$$ ∎

4. Error estimation procedure. We apply Algorithm 1 to the example from section 2. To ensure global convergence, error estimates of the form (3.15) have to be fulfilled. We discuss how they can be ensured for the objective function (2.2) and the state equation (2.1). The index k denoting the iteration number in the algorithm is skipped for readability purposes in this section as far as possible.

Model gradient error. The model gradient is computed by the adjoint approach with inexact solutions $\tilde{y}$ and $\tilde{z}$ of the state and the adjoint equation, respectively. Let $y = S(u) \in Y$ be the exact state solving $E(S(u),u) = 0$ and let $\tilde{y} \in Y$ be an inexact solution. The perturbed adjoint equation reads

$$E_y(\tilde{y},u)^* \hat{z} = -J_y(\tilde{y},u). \tag{4.1}$$

Its exact solution is denoted by $\hat{z}$ and its inexact solution by $\tilde{z}$, whereas z is the exact adjoint state solving

$$E_y(y,u)^* z = -J_y(y,u). \tag{4.2}$$

THEOREM 4.1. *Let $J : Y \times U \to \mathbb{R}$ be defined as in (2.2), let $E : Y \times U \to Y^*$ be as in (2.1), and let $u \in U$ and $\tilde{y}, \tilde{z} \in Y$ be given. Moreover, let $y \in Y$ be the exact solution of $E(y,u) = 0$ and let $z \in Y$ and $\hat{z} \in Y$ be the exact solutions of (4.2) and (4.1), respectively.*
Choosing m such that $m'(0) = E_u(\tilde{y},u)^* \tilde{z} + J_u(\tilde{y},u)$, it then holds that

$$\begin{aligned} \|m'(0) - \hat{J}'(u)\|_{U^*} \le{}& \|B^*(\hat{z}-\tilde{z})\|_{U^*} + \tfrac{1}{\underline{\kappa}} \|B\|_{\mathscr{L}(U,Y^*)} \|Q\|_{\mathscr{L}(Y,H)} \big(\|Q(y-\tilde{y})\|_H \\ &+ \tfrac{1}{\underline{\kappa}} \|N'(\tilde{y}) - N'(y)\|_{\mathscr{L}(Y,Y^*)} \big(\|Q\tilde{y} - \hat{q}\|_H + \|Q(y-\tilde{y})\|_H \big) \big). \end{aligned} \tag{4.3}$$

Proof. From the choice of $m'(0)$ and (4.2) we obtain using that $E_u \equiv -B$ and that J_u is independent of y:

$$\begin{aligned} \|m'(0) - \hat{J}'(u)\|_{U^*} &= \|E_u(\tilde{y},u)^* \tilde{z} - E_u(y,u)^* z + J_u(\tilde{y},u) - J_u(y,u)\|_{U^*} \\ &= \|B^*(z-\tilde{z})\|_{U^*} \le \|B^*(z-\hat{z})\|_{U^*} + \|B^*(\hat{z}-\tilde{z})\|_{U^*}. \end{aligned} \tag{4.4}$$

The monotonicity of N yields the monotonicity of $N'(y)$ for every $y \in Y$ and thus the strong monotonicity of $A^* + N'(y)^*$. Therefore, both equations (4.1) and (4.2) are uniquely solvable and we can compute

$$\begin{aligned} \hat{z} - z = {}& -E_y(\tilde{y},u)^{-*} J_y(\tilde{y},u) + E_y(y,u)^{-*} J_y(y,u) \\ ={}& E_y(\tilde{y},u)^{-*} (J_y(y,u) - J_y(\tilde{y},u)) + (E_y(y,u)^{-*} - E_y(\tilde{y},u)^{-*}) J_y(\tilde{y},u) \\ &+ (E_y(y,u)^{-*} - E_y(\tilde{y},u)^{-*})(J_y(y,u) - J_y(\tilde{y},u)). \end{aligned} \tag{4.5}$$

With (2.2) and (2.1), the perturbed adjoint equation (4.1) reads

$$E_y(\tilde{y},u)^* z = A^* z + N'(\tilde{y})^* z = -Q^*(Q\tilde{y} - \hat{q}) = -J_y(\tilde{y},u). \tag{4.6}$$

Compared to the exact adjoint equation (4.2), the error in the right-hand side is

$$J_y(y,u) - J_y(\tilde{y},u) = Q^* Q(y - \tilde{y}). \tag{4.7}$$

For the estimation of the error caused by the approximate left-hand side operator, we introduce for $\tilde{b} \in Y^*$ the unique solutions $v, \tilde{v} \in Y$ of the equations

$$A^* v + N'(y)^* v = \tilde{b}, \qquad A^* \tilde{v} + N'(\tilde{y})^* \tilde{v} = \tilde{b},$$

respectively. We have $v-\tilde{v} = (E_y(y,u)^{-*} - E_y(\tilde{y},u)^{-*})\tilde{b}$. Using the monotonicity of $N'(y)$ and the strong monotonicity of A with constant $\underline{\kappa}$, we estimate:

$$\begin{aligned}\underline{\kappa}\|v-\tilde{v}\|_Y^2 &\le \langle v-\tilde{v}, A(v-\tilde{v})\rangle_{Y,Y^*} \le \langle v-\tilde{v}, (A+N'(y))(v-\tilde{v})\rangle_{Y,Y^*}\\ &= \langle A^*v - A^*\tilde{v} + N'(y)^*v - N'(y)^*\tilde{v}, v-\tilde{v}\rangle_{Y^*,Y}\\ &= \langle \tilde{b} - A^*\tilde{v} - N'(\tilde{y})^*\tilde{v} + N'(\tilde{y})^*\tilde{v} - N'(y)^*\tilde{v}, v-\tilde{v}\rangle_{Y^*,Y}\\ &= \langle (N'(\tilde{y}) - N'(y))^*\tilde{v}, v-\tilde{v}\rangle_{Y^*,Y} \le \|(N'(\tilde{y}) - N'(y))^*\tilde{v}\|_{Y^*}\|v-\tilde{v}\|_Y.\end{aligned}$$

This results in

$$\begin{aligned}\|(E_y(y,u)^{-*} - E_y(\tilde{y},u)^{-*})\tilde{b}\|_Y &\le \tfrac{1}{\underline{\kappa}}\|(N'(\tilde{y}) - N'(y))^*\tilde{v}\|_{Y^*}\\ &\le \tfrac{1}{\underline{\kappa}^2}\|N'(\tilde{y}) - N'(y)\|_{\mathscr{L}(Y,Y^*)}\|\tilde{b}\|_{Y^*},\end{aligned} \tag{4.8}$$

where the last estimate is due to the strong monotonicity of $A^* + N'(\tilde{y})^*$. Inserting (4.7) into (4.5) and using (4.8), we obtain (again using strong monotonicity):

$$\begin{aligned}\|\hat{z}-z\|_Y \le{}& \tfrac{1}{\underline{\kappa}}\|Q^*Q(y-\tilde{y})\|_{Y^*}\\ &+ \tfrac{1}{\underline{\kappa}^2}\|N'(\tilde{y}) - N'(y)\|_{\mathscr{L}(Y,Y^*)}\big(\|Q^*(Q\tilde{y}-\hat{q})\|_{Y^*} + \|Q^*Q(y-\tilde{y})\|_{Y^*}\big).\end{aligned} \tag{4.9}$$

Combining this and (4.4) results in (4.3). ∎

To bound the gradient error, we therefore have to control the error $\|B^*(\hat{z}-\tilde{z})\|_{U^*}$ caused by the inexact solution of the perturbed adjoint equation (4.1). If, e.g., $B \equiv \iota : L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$, it is sufficient to control the adjoint state error in the $L^2(\Omega)$-norm. This is no longer true if a boundary control problem is considered. We will therefore estimate $\|B^*(\hat{z}-\tilde{z})\|_{U^*} \le \|B\|_{\mathscr{L}(U,Y^*)}\|\hat{z}-\tilde{z}\|_Y$ and control the error in the Y-norm. Another reason for this is that a posteriori error estimation techniques to estimate the $L^2(\Omega)$-error require the PDE solution to have $H^2(\Omega)$-regularity, see, e.g., [1, sec. 2.4]. This cannot be guaranteed if the coefficient function κ in the definition (2.6) of the operator A is only $L^\infty(\Omega)$-regular, i.e., it can contain jumps along edges for example, or if the domain Ω is non-convex.

Moreover, we have to control the errors $\|Q(y-\tilde{y})\|_H$ or even $\|Q^*Q(y-\tilde{y})\|_{Y^*}$, see (4.9), and $\|N'(\tilde{y}) - N'(y)\|_{\mathscr{L}(Y,Y^*)}$ introduced by the inexact solution of the state equation. Again, if, e.g., $Q \equiv \iota : H_0^1(\Omega) \hookrightarrow L^2(\Omega)$, it is sufficient to control the state error in the $L^2(\Omega)$-norm, which is no longer true if we have a problem with, e.g., boundary observation. If N' is locally Lipschitz continuous w.r.t. y, we can bound $\|N'(\tilde{y}) - N'(y)\|_{\mathscr{L}(Y,Y^*)} \le c_{N'}\|\tilde{y}-y\|_Y$. Then we need the state error in the stronger Y-norm. For the example from subsection 2.1, the local Lipschitz constant can be bounded as follows:

LEMMA 4.2. *Let $N : Y \to Y^*$ be defined as in (2.6) and let $\varphi : \mathbb{R} \to \mathbb{R}$ be increasing, twice continuously differentiable, and fulfilling (2.5). Then we have that*

$$\|N'(\tilde{y}) - N'(y)\|_{\mathscr{L}(Y,Y^*)} \le c_p^3\Big(a''_\varphi\,\lambda(\Omega)^{(p-3)/p} + c''_\varphi c_p^{p-3}\left(\|\tilde{y}\|_Y + \|y-\tilde{y}\|_Y\right)^{p-3}\Big)\|y-\tilde{y}\|_Y,$$

where λ is the Lebesgue measure on Ω and $c_p > 0$ is the Sobolev constant such that $\|y\|_{L^p(\Omega)} \le c_p\|y\|_{H_0^1(\Omega)}$ holds for every $y \in Y$.

Proof. We have that $\langle N'(y)v, \tilde{v}\rangle_{Y^*,Y} = \int_\Omega \varphi'(y)v\tilde{v}\,\mathrm{d}x$ for $y, v, \tilde{v} \in Y$. Thus, using (2.5), we can estimate with $r_i \in [1,\infty]$ ($i \in \{1,2,3,4,5\}$), $\frac{1}{r_1}+\frac{1}{r_2}+\frac{1}{r_3} = 1$, $\frac{1}{r_4}+\frac{1}{r_5} = \frac{1}{r_1}$, and $r_4(p-3) \ge 1$

(to be specified later), and writing $\tilde{p} = p-3$:

$$\begin{aligned}
&|\langle (N'(y)-N'(\tilde{y}))v,\tilde{v}\rangle_{Y^*,Y}| \le \|\varphi'(y)-\varphi'(\tilde{y})\|_{L^{r_1}(\Omega)}\|v\|_{L^{r_2}(\Omega)}\|\tilde{v}\|_{L^{r_3}(\Omega)}\\
&\le \int_0^1 \|\varphi''(\tilde{y}+\tau(y-\tilde{y}))(y-\tilde{y})\|_{L^{r_1}(\Omega)}\,\mathrm{d}\tau\,\|v\|_{L^{r_2}(\Omega)}\|\tilde{v}\|_{L^{r_3}(\Omega)}\\
&\le \Big(a''_\varphi\,\lambda(\Omega)^{1/r_4} + c''_\varphi \sup_{\tau\in[0,1]} \|\tilde{y}+\tau(y-\tilde{y})\|^{\tilde{p}}_{L^{r_4\tilde{p}}(\Omega)}\Big)\|y-\tilde{y}\|_{L^{r_5}(\Omega)}\|v\|_{L^{r_2}(\Omega)}\|\tilde{v}\|_{L^{r_3}(\Omega)}\\
&\le c_{r_2}c_{r_3}\Big(a''_\varphi\,\lambda(\Omega)^{1/r_4} + c''_\varphi\left(\max\{\|y\|_{L^{r_4\tilde{p}}(\Omega)},\|\tilde{y}\|_{L^{r_4\tilde{p}}(\Omega)}\}\right)^{\tilde{p}}\Big)\|y-\tilde{y}\|_{L^{r_5}(\Omega)}\|v\|_Y\|\tilde{v}\|_Y,
\end{aligned}$$

where $c_{\hat{r}}$ is the constant from the Sobolev embedding $Y = H^1_0(\Omega) \hookrightarrow L^{\hat{r}}(\Omega)$ with adequately chosen $\hat{r}\in[1,\infty)$ or $\hat{r}\in[1,6]$ dependent on n. Choosing $r_2 = r_3 = \tilde{r}$ with $\tilde{r}\in(2,\infty)$ for $n=2$ and $\tilde{r}\in(2,6]$ for $n=3$, $r_5 = \frac{\tilde{r}(p-2)}{\tilde{r}-2} = p-2+\frac{2(p-2)}{\tilde{r}-2} > p-2 > 1$, and $r_4 = \frac{r_5}{\tilde{p}} > 1+\frac{1}{p-3}$, this gives

$$\|N'(y)-N'(\tilde{y})\|_{\mathscr{L}(Y,Y^*)} \le c_{\tilde{r}}^2\Big(a''_\varphi\,\lambda(\Omega)^{\tilde{p}/r_5} + c''_\varphi\left(\max\{\|y\|_{L^{r_5}(\Omega)},\|\tilde{y}\|_{L^{r_5}(\Omega)}\}\right)^{\tilde{p}}\Big)\|y-\tilde{y}\|_{L^{r_5}(\Omega)}.$$

The concrete choice $\tilde{r} = p$ (giving $r_5 = p$, $r_4 = \frac{p}{p-3}$) and the Sobolev embedding $Y \hookrightarrow L^p(\Omega)$ yield the result stated in the lemma. ∎

Projection error. In order to successfully apply Proposition 3.4 and to implemented the projected linesearch, a projection error bound (3.15) (c) is needed. In many cases, such as a finite dimensional control space $U = \mathbb{R}^{n_u}$ and a simple feasible set (e.g., a box), or the space $U = L^2(\Omega_u)$ with Ω_u being a measurable subset of Ω or $\partial\Omega$ and a norm ball constraint, the projection can be computed exactly. The same holds true for $U = L^2(\Omega_u)$ and $U_{\mathrm{ad}} := \{u\in U : u_{\mathrm{l}} \le u \le u_{\mathrm{u}} \text{ a.e.}\}$ with $u_{\mathrm{l}} < u_{\mathrm{u}} \in \mathbb{R}$ and a discretization by piecewise constant functions. In contrast, if piecewise linear FE functions are used, one would like to compute an approximate projection by pointwisely projecting the nodal function values onto the box. Then, an error occurs on the elements, where the function crosses a bound. This element-wise error can be computed exactly [22] and can be reduced by refining the elements with the largest error contribution. In our implementation, we pick as many triangles as needed to cover a certain amount $\vartheta_P \in (0,1)$ of the total error, more concretely $\vartheta_P = 30\%$.

Objective function evaluation error. We assume that the inexact reduced objective function $\hat{J}_k$ is evaluated using an inexact solution $\tilde{y}\in Y$ of the state equation, i.e., $\hat{J}(u) = J(S(u),u)$ and $\hat{J}_k(u) = J(\tilde{y},u)$ for some $\tilde{y}\in Y$. To derive (3.15) (a), we have to bound $|\hat{J}(u)-\hat{J}_k(u)| = |J(y,u)-J(\tilde{y},u)|$ for $u\in U$, where $y = S(u)$ is the exact solution of the state equation. If J is locally Lipschitz continuous w.r.t. y, it holds that $|J(y,u)-J(\tilde{y},u)| \le c_J\|y-\tilde{y}\|_Y$, and for error estimation we have to estimate the error in the computed state and possibly the local Lipschitz constant c_J. For a tracking-type objective function, we have a more explicit estimate:

PROPOSITION 4.3. *Let $J: Y\times U\to\mathbb{R}$ be of tracking type form (2.2) and let $y,\tilde{y}\in Y$, $u\in U$ be given. Then the following estimate holds true:*

$$|J(y,u)-J(\tilde{y},u)| \le \tfrac{1}{2}\|Q(y-\tilde{y})\|_H^2 + \|Q\tilde{y}-\hat{q}\|_H\|Q(y-\tilde{y})\|_H. \tag{4.10}$$

Proof. The estimate (4.10) follows from

$$\begin{aligned}
J(y,u)-J(\tilde{y},u) &= \tfrac{1}{2}\|Qy-Q\tilde{y}+Q\tilde{y}-\hat{q}\|_H^2 - \tfrac{1}{2}\|Q\tilde{y}-\hat{q}\|_H^2\\
&= \tfrac{1}{2}\|Q(y-\tilde{y})\|_H^2 + (Q(y-\tilde{y}),Q\tilde{y}-\hat{q})_H.
\end{aligned}$$

∎

We see that we have to estimate the error $\|Q(y-\tilde{y})\|_H \leq \|Q\|_{\mathscr{L}(Y,H)}\|y-\tilde{y}\|_Y$, which again reduces to the question of state error estimation. An alternative to the estimate given in Proposition 4.3 would be the dual-weighted-residual method [2], which is well-suited for the estimation of the error in the objective function of an optimal control problem. We do not employ it here having the stochastic application [12] in mind. We want to rely on already established error estimation techniques for PDEs with uncertain inputs, which can be implemented with low-rank tensors.

In conclusion, we have to control the errors in the inexact state and adjoint state as well as the error caused by the inexact projection to ensure (3.15).

5. Adaptive solution of the PDEs. We now discuss the discretization of the problem, in particular of the underlying equations, and the a posteriori error estimation. We use conforming finite element discretizations for the deterministic state and control spaces. The discretization will be adaptive, i.e., we have sequences of nested discrete spaces with their respective bases and linear maps prolongating the coefficients from coarser to finer spaces. We only describe the discretization on a fixed grid, having in mind that the mesh size will be adapted in the final implementation. In the following, we assume that the domain $\Omega \subset \mathbb{R}^2$ is polygonal and that the restriction $D|_{\mathrm{U}} : \mathrm{U} \to L^2(\Omega)$ of the control operator D to any used discrete subspace $\mathrm{U} \subset U$ can be evaluated exactly. We exclude the case $\Omega \subset \mathbb{R}^3$ here, which leads to a more compact presentation of a posteriori error estimates in subsection 5.2. It is possible to generalize many results to the 3D case.

5.1. Space discretization. The domain Ω is partitioned into a finite element mesh yielding a triangulation $\mathscr{T}$. Let $\mathrm{Y} \subset Y$ denote the discrete subspace of piecewise linear, globally continuous finite element functions with zero boundary data. The nodal FE basis of the discrete, deterministic state space is $\{\phi_{k_0}\}_{k_0=1}^{d_0} \subset \mathrm{Y} \subset Y = H_0^1(\Omega)$ and the basis of thr discrete control space $\mathrm{U} \subset U = L^2(\Omega_u)$ is denoted by $\{\psi_\ell\}_{\ell=1}^{d_u}$. This can also be a nodal FE basis if Ω_u is a subset of Ω with positive measure or—for finite-dimensional controls—the standard basis of $\mathbb{R}^{d_u} = L^2([d_u])$. In the latter case, no discretization is required. The basis functions shall sum to one, i.e., $\sum_{k_0=1}^{d_0} \phi_{k_0}(x) = 1$ and $\sum_{k_u=1}^{d_u} \psi_{k_u}(\tilde{x}) = 1$ for all $x \in \Omega$, $\tilde{x} \in \Omega_u$ to perform mass lumping in a meaningful way later.

In addition to assuming that $D|_{\mathrm{U}} : L^2(\Omega_u) \to L^2(\Omega)$ can be evaluated exactly, we match the discretizations of the state and the control space, i.e., we use the same grid on Ω and Ω_u if Ω_u is a subset of Ω with positive measure. This makes the computation of the gradient of the reduced objective function easier since $\hat{J}'(u) = B^*z + \gamma(u,\cdot)_U$.

We define the following matrices:

- the mass matrix $\mathsf{M} \in \mathbb{R}^{d_0 \times d_0}$ for Y: $\mathsf{M}_{k_0 l_0} := (\phi_{k_0}, \phi_{l_0})_{L^2(\Omega)}$,
- the lumped mass matrix $\mathsf{M}_\mathsf{L} \in \mathbb{R}^{d_0 \times d_0}$ for Y: $(\mathsf{M}_\mathsf{L})_{k_0 k_0} := \sum_{l_0=1}^{d_0} \mathsf{M}_{k_0 l_0} = \int_\Omega \phi_{k_0}\,\mathrm{d}x$ and $(\mathsf{M}_\mathsf{L})_{k_0 l_0} = 0$ for $k_0 \neq l_0$,
- the mass matrix $\tilde{\mathsf{M}} \in \mathbb{R}^{d_u \times d_u}$ for U: $\tilde{\mathsf{M}}_{k_u l_u} := (\psi_{k_u}, \psi_{l_u})_{L^2(\Omega_u)}$.

Let $\mathsf{y} \in \mathbb{R}^{d_0}$ and $\mathsf{u} \in \mathbb{R}^{d_u}$ be the coefficients, representing the discrete state y and control u, respectively. Inserting $\mathrm{y}(x) = \sum_{k_0=1}^{d_0} \mathsf{y}_{k_0}\phi_{k_0}(x)$ and $\mathrm{u}(x) = \sum_{k_u=1}^{d_u} \mathsf{u}_{k_u}\psi_{k_u}(x)$ into (2.1) with the definition (2.6), and testing with $v \equiv \phi_{k_0}$ for $k_0 \in [d_0]$, the discrete version of the deterministic state equation reads

$$
\begin{aligned}
&\mathsf{A}\mathsf{y} + \mathsf{N}(\mathsf{y}) = \mathsf{B}\mathsf{u} + \mathsf{b}, \\
&\mathsf{A} \in \mathbb{R}^{d_0 \times d_0}, \mathsf{A}_{k_0 l_0} = (\kappa \nabla\phi_{l_0}, \nabla\phi_{k_0})_{L^2(\Omega)^n}, \quad && \mathsf{N} : \mathbb{R}^{d_0} \to \mathbb{R}^{d_0}, \mathsf{N}(\mathsf{y}) = \mathsf{M}_\mathsf{L}\varphi(\mathsf{y}), \\
&\mathsf{B} \in \mathbb{R}^{d_0 \times d_u}, \mathsf{B}_{k_0 l_u} = (D\psi_{l_u}, \phi_{k_0})_{L^2(\Omega)}, && \mathsf{b} \in \mathbb{R}^{d_0}, \mathsf{b}_{k_0} = (f, \phi_{k_0})_{L^2(\Omega)}.
\end{aligned} \tag{5.1}
$$

Due to the ease of implementation and interpolation, a quadrature error is allowed to occur

in the discretization of the nonlinearity which is connected to mass lumping: The integral $\int_\Omega \varphi(y)\phi_{k_0}\,dx$ is evaluated inexactly by a quadrature formula, the nodes of which are the finite element grid nodes and the weights of which are the respective entries of the lumped mass matrix. We obtain

$$\int_\Omega \varphi(y)\phi_{k_0}\,dx \approx \varphi(y_{k_0})(M_L)_{k_0k_0}. \tag{5.2}$$

Since Algorithm 1 is formulated in the space U, it is desirable to not make additional errors by mass lumping in the objective function, but to evaluate the U- and H-inner product exactly. Let H be discretized such that $Q|_Y$ can be evaluated exactly and $\hat{q}$ can be represented exactly. The discrete subspace H of H is isomorphic to $\mathbb{R}^{d_H}$ ($d_H \in \mathbb{N}$) equipped with the inner product induced by the symmetric, positive definite matrix $M_H \in \mathbb{R}^{d_H\times d_H}$. Let $Q \in \mathbb{R}^{d_H\times d_0}$ and $\hat{q} \in \mathbb{R}^{d_H}$ be the discrete versions of Q and $\hat{q}$, respectively. Then, the discretized objective function reads

$$J(y,u) = \tfrac{1}{2}\|Qy-\hat{q}\|^2_{M_H} + \tfrac{\gamma}{2}u^\top \tilde{M}u, \tag{5.3}$$

cf. (2.2). We note that under the stated assumptions, the evaluation of the objective function is exact so that the error in the reduced objective function depends only on the error in the discretized state and Proposition 4.3 can be applied. The discrete version of the deterministic adjoint equation (2.8) is

$$Az + M_L(\varphi'(y)\odot z) = -Q^\top M_H(Qy-\hat{q}), \tag{5.4}$$

where again the quadrature formula using the finite element nodes has been applied. In fact, $N'(y)z = M_L(\varphi'(y)\odot z)$ holds also in the discrete setting and we can identify $N'(y) = M_L\,\mathrm{diag}(\varphi'(y))$. The gradient of the reduced, deterministic, discretized objective function is then given by

$$\nabla \hat{J}(u) = -\tilde{M}^{-1}B^\top z + \gamma u, \tag{5.5}$$

cf. (2.9). Note that in typical situations it is not necessary to invert the mass matrix $\tilde{M}$ to compute the reduced gradient: If, e.g., $\Omega_u \subset \Omega$ is a subset of positive measure, $B^* : H^1_0(\Omega) \to L^2(\Omega_u), z \mapsto z|_{\Omega_u}$ is the canonical embedding $\iota : H^1_0(\Omega) \hookrightarrow L^2(\Omega)$ combined with restriction of the function to Ω_u, and the grids on Ω and Ω_u match, the application of $\tilde{M}^{-1}B$ consists of a simple extraction of components of the vector z and/or adding zero components for the nodes on $\partial\Omega$. Such situations are favorable because the error in the discrete gradient then only depends on the error in the discrete adjoint state and it is sufficient to apply the error estimate from Theorem 4.1.

Once the equations (2.12) and (2.11) are discretized, the application of the Hessian operator to a direction can be computed via (2.10). For this purpose, it only remains to discretize the term in (2.11) involving the second derivative N'', which is again done by using the FE nodes based quadrature. Then, if $s \in \mathbb{R}^{d_u}$ represents a direction $s \in U$, and y, z represent the current state and adjoint state, respectively, the application of the Hessian to this direction reads

$$\nabla^2 \hat{J}(u)s = \tilde{M}^{-1}B^\top h + \gamma s, \tag{5.6}$$

where h solves

$$[A+N'(y)]h = Q^\top M_H Qd + M_L\big(z\odot\varphi''(y)\odot d\big) \quad \text{with } d = [A+N'(y)]^{-1}Bs.$$

Choice of the trust-region model. Let $\mathrm{U} \subset U$ be the current discrete subspace and let $\mathsf{u}^k \in \mathrm{U}$ be the current iterate, represented by the vector $\mathsf{u}^k \in \mathbb{R}^{d_u}$. The vectors $\tilde{\mathsf{y}}^k$ and $\tilde{\mathsf{z}}^k$ shall (approximately) solve the state equation (5.1) with $\mathsf{u} = \mathsf{u}^k$ and the adjoint equation (5.4) with $\mathsf{y} = \tilde{\mathsf{y}}^k$, respectively. In order to be able to apply the error estimates from section 4, we choose a quadratic trust-region model m_k for any direction $\mathrm{s} \in \mathrm{U}$ represented by the vector s as follows:

$$m_k(\mathrm{s}) = \mathsf{m}_k(\mathsf{s}) := \nabla \mathsf{m}_k(0)^\top \tilde{\mathsf{M}}\mathsf{s} + \tfrac{1}{2}\mathsf{s}^\top \tilde{\mathsf{M}} \nabla^2 \mathsf{m}_k(0)\mathsf{s} \tag{5.7}$$

The space $\mathbb{R}^{d_u}$ is equipped with the inner product induced by the mass matrix $\tilde{\mathsf{M}}$ and derivatives, such as $\nabla \mathsf{m}_k$, are computed w.r.t. this inner product. To approximate the true gradient sufficiently well, we choose

$$\nabla \mathsf{m}_k(0) = -(\tilde{\mathsf{M}}^{-1}\mathsf{B}^\top)\tilde{\mathsf{z}}^k + \gamma \mathsf{u}^k \tag{5.8}$$

in light of (5.5) and Theorem 4.1. In the deterministic case, we choose $\nabla^2 m_k(0)$ to be computed as in (5.6) using $\tilde{\mathsf{y}}^k$ and $\tilde{\mathsf{z}}^k$, although we would have more freedom to approximate the true Hessian. This is used in the stochastic case [12].

5.2. A posteriori error estimation. Following [3, 4, 7, 8, 9] we consider the elliptic, nonlinear operator equation

$$\hat{A}y + \hat{N}(y) = \hat{b}. \tag{5.9}$$

on the Hilbert space Y with $\hat{b} \in Y^*$. Let $\hat{N} : Y \to Y^*$ be well-defined, continuous and monotone, but possibly nonlinear, and let $\hat{A} \in \mathscr{L}(Y, Y^*)$ be self-adjoint and boundedly invertible:

$$\|\hat{A}\|_{\mathscr{L}(Y,Y^*)} \le c_{\max}, \quad \|\hat{A}^{-1}\|_{\mathscr{L}(Y^*,Y)} \le c_{\min}^{-1}.$$

This gives rise to the inner product $(y,v)_{\hat{A}} := \langle \hat{A}y, v\rangle_{Y^*,Y}$ and the corresponding energy norm $\|y\|_{\hat{A}} = \sqrt{(y,y)_{\hat{A}}}$.

For error estimation, we consider a linear, self-adjoint and elliptic "reference" operator $\hat{A}_{\mathrm{ref}} \in \mathscr{L}(Y, Y^*)$. In particular, this can be a deterministic nominal operator whereas $\hat{A}$ is uncertain, see [12], or any other norm-inducing operator, such as $\hat{A}_{\mathrm{ref}} = -\Delta$ for the $H_0^1(\Omega)$-norm. Then, $(y,v)_{\hat{A}_{\mathrm{ref}}} := \langle \hat{A}_{\mathrm{ref}} y, v\rangle_{Y^*,Y}$ defines an alternative inner product on the space Y and the equivalence estimate

$$\lambda \langle \hat{A}v, v\rangle_{Y^*,Y} \le \langle \hat{A}_{\mathrm{ref}} v, v\rangle_{Y^*,Y} \le \Lambda \langle \hat{A}v, v\rangle_{Y^*,Y} \tag{5.10}$$

holds for every $v \in Y$ with some constants $\lambda, \Lambda \in (0,\infty)$, $\lambda \le \Lambda$. We define the alternative norm $\|y\|_{\hat{A}_{\mathrm{ref}}} := \sqrt{\langle \hat{A}_{\mathrm{ref}} y, y\rangle_{Y^*,Y}}$ on Y which is equivalent to the usual norm $\|\cdot\|_Y$. Analogously defining the inner products and norms on Y^* induced by the operators $\hat{A}^{-1}$ and $\hat{A}_{\mathrm{ref}}^{-1}$, one can show the inverse estimates

$$\tfrac{1}{\Lambda}\|\tilde{b}\|_{\hat{A}^{-1}}^2 \le \langle \tilde{b}, \hat{A}_{\mathrm{ref}}^{-1}\tilde{b}\rangle_{Y^*,Y} = \|\tilde{b}\|_{\hat{A}_{\mathrm{ref}}^{-1}}^2 \le \tfrac{1}{\lambda}\|\tilde{b}\|_{\hat{A}^{-1}}^2 \text{ for every } \tilde{b} \in Y^*. \tag{5.11}$$

Now, let $y \in Y$ be the unique and exact solution of (5.9), and let $\tilde{\mathrm{y}} \in \mathrm{Y} \subset Y$ be an inexact solution living in a subspace Y of Y (e.g., a finite element subspace), fulfilling

$$\hat{A}\tilde{\mathrm{y}} + \hat{N}(\tilde{\mathrm{y}}) - \hat{b} =: r, \tag{5.12}$$

where $r \in Y^*$ is the residual. We assume that that this residual can be evaluated exactly in the sense that $\langle r, \mathrm{v}^+\rangle_{Y^*,Y}$ can be evaluated for every given $\mathrm{v}^+ \in \mathrm{Y}^+ \supset \mathrm{Y}$ in some finite-dimensional (e.g., FE) subspace $\mathrm{Y}^+ \subset Y$.

LEMMA 5.1. *Let $y \in Y$ be the solution of (5.9) and let $\tilde{y} \in Y$. Under the standing assumptions, and with r defined in (5.12), we have*

$$\|\tilde{y}-y\|_{\hat{A}_{\mathrm{ref}}} \le \Lambda \|r\|_{\hat{A}_{\mathrm{ref}}^{-1}}. \tag{5.13}$$

Proof. We can estimate using the monotonicity of $\hat{N}$:

$$\begin{aligned}\|\tilde{y}-y\|_{\hat{A}}^2 \le& \langle \hat{A}(\tilde{y}-y), \tilde{y}-y\rangle_{Y^*,Y} + \langle \hat{N}(\tilde{y})-\hat{N}(y), \tilde{y}-y\rangle_{Y^*,Y} = \langle r, \tilde{y}-y\rangle_{Y^*,Y}\\ =& \langle \hat{A}\hat{A}^{-1}r, \tilde{y}-y\rangle_{Y^*,Y} = (\hat{A}^{-1}r, \tilde{y}-y)_{\hat{A}} \le \|\hat{A}^{-1}r\|_{\hat{A}}\|\tilde{y}-y\|_{\hat{A}}.\end{aligned}$$

This gives $\|\tilde{y}-y\|_{\hat{A}} \le \|r\|_{\hat{A}^{-1}}$ and

$$\frac{1}{\sqrt{\Lambda}}\|\tilde{y}-y\|_{\hat{A}_{\mathrm{ref}}} \le \|\tilde{y}-y\|_{\hat{A}} \le \|r\|_{\hat{A}^{-1}} \le \sqrt{\Lambda}\|r\|_{\hat{A}_{\mathrm{ref}}^{-1}} \tag{5.14}$$

by the norm estimates (5.10) and (5.11), which yields the result. ∎

For the computation of $\|r\|_{\hat{A}_{\mathrm{ref}}^{-1}}$ we define $w \in Y$ to be the unique solution of the equation $\hat{A}_{\mathrm{ref}} w = r$. Then it holds that

$$\|r\|_{\hat{A}_{\mathrm{ref}}^{-1}}^2 = \langle r, \hat{A}_{\mathrm{ref}}^{-1} r\rangle_{Y^*,Y} = \langle \hat{A}_{\mathrm{ref}} w, w\rangle_{Y^*,Y} = \|w\|_{\hat{A}_{\mathrm{ref}}}^2. \tag{5.15}$$

We compute a discrete solution $\mathrm{w} \in \mathrm{Y}$ fulfilling

$$\langle \hat{A}_{\mathrm{ref}}\mathrm{w}, \mathrm{v}\rangle_{Y^*,Y} = \langle r, \mathrm{v}\rangle_{Y^*,Y} \text{ for all } \mathrm{v} \in \mathrm{Y}, \tag{5.16}$$

i.e., $\mathrm{w} = \hat{\mathrm{A}}_{\mathrm{ref}}^{-1} r$, where $\hat{\mathrm{A}}_{\mathrm{ref}} : \mathrm{Y} \to Y^*$ is the restriction of the operator $\hat{A}_{\mathrm{ref}}$ onto the space Y and its inverse is defined in the sense of (5.16). We assume that this equation is solved exactly having our application in mind. If this it not the case, the algebraic error caused by the inexact solution of the discrete equation has to be incorporated additionally.

LEMMA 5.2. *Let $y, \tilde{y}, r$ be as in Lemma 5.1, and let $w = \hat{A}_{\mathrm{ref}}^{-1} r$ and $\mathrm{w} \in \mathrm{Y}$ defined by (5.16). Then, under the standing assumptions, we have that*

$$\|\tilde{y}-y\|_{\hat{A}_{\mathrm{ref}}}^2 \le \Lambda^2\big(\|\mathrm{w}\|_{\hat{A}_{\mathrm{ref}}}^2 + \|w-\mathrm{w}\|_{\hat{A}_{\mathrm{ref}}}^2\big). \tag{5.17}$$

Proof. Combining (5.13) and (5.15) results in

$$\|\tilde{y}-y\|_{\hat{A}_{\mathrm{ref}}}^2 \le \Lambda^2\big(\|\mathrm{w}+w-\mathrm{w}\|_{\hat{A}_{\mathrm{ref}}}^2\big) = \Lambda^2\big(\|\mathrm{w}\|_{\hat{A}_{\mathrm{ref}}}^2 + \|w-\mathrm{w}\|_{\hat{A}_{\mathrm{ref}}}^2\big), \tag{5.18}$$

where the last equality is due to (5.16) and $\mathrm{w} \in \mathrm{Y}$, cf. the proof of [7, Thm. 5.1]. ∎

The first summand in (5.17) turns out to be the purely algebraic error contribution caused by solving a discretized version of (5.9) inexactly:

$$\|\mathrm{w}\|_{\hat{A}_{\mathrm{ref}}}^2 = \langle r, \hat{\mathrm{A}}_{\mathrm{ref}}^{-1} r\rangle_{Y^*,Y} = \langle \hat{A}\tilde{y} + \hat{N}(\tilde{y}) - \hat{b}, \hat{\mathrm{A}}_{\mathrm{ref}}^{-1}(\tilde{y} + \hat{N}(\tilde{y}) - \hat{b})\rangle_{Y^*,Y}. \tag{5.19}$$

The second summand will be estimated by a posteriori error estimates for (5.16).

Realization of the a posteriori error estimator. We discuss the realization of a deterministic a posteriori error estimator for the estimation of the term $\|w-\mathrm{w}\|_{\hat{A}_{\mathrm{ref}}}$ for the example from subsection 2.1 with the deterministic state equation (2.1) and the adjoint equation (2.8). This is an adaption of the ideas presented in [20, chap. 1] and [1, chap. 2] to our setting.

We define in analogy to (2.6), but with slight differences:

$$\begin{aligned}
&\langle \hat{A}y, v\rangle_{Y^*,Y} := \int_\Omega \hat{\kappa}\nabla y\cdot\nabla v + \hat{\chi}yv\,\mathrm{d}x, \quad &&\langle \hat{N}(y), v\rangle_{Y^*,Y} := \int_\Omega \hat{\varphi}(y)v\,\mathrm{d}x,\\
&\langle \hat{b}, v\rangle_{Y^*,Y} := \int_\Omega \hat{f}v\,\mathrm{d}x, &&\langle \hat{A}_{\mathrm{ref}}y, v\rangle_{Y^*,Y} := \int_\Omega \hat{\kappa}_{\mathrm{ref}}\nabla y\cdot\nabla v + \hat{\chi}_{\mathrm{ref}}yv\,\mathrm{d}x
\end{aligned} \tag{5.20}$$

with $\hat{\chi}\in L^\infty(\Omega)$, $\hat{\chi}(x)\geq 0$ for a.e. $x\in\Omega$ and the reference coefficients $\hat{\kappa}_{\mathrm{ref}}$ (uniformly positive) and $\hat{\chi}_{\mathrm{ref}}$ (nonnegative). The function $\hat{\varphi}\in\mathscr{C}^2(\mathbb{R},\mathbb{R})$ shall also be increasing and fulfill (2.5).

Remark 5.3. The setting covers all important equations and operators:

- For $\hat{\chi}\equiv 0$, $\hat{\varphi}\equiv\varphi$, and $\hat{f}=Du+f$ we get the state equation (2.1) with the operators defined in (2.6).
- If, e.g., $Q\equiv\iota : H_0^1(\Omega)\hookrightarrow L^2(\Omega)$, we get the adjoint equation (2.8) for $\hat{\chi}\equiv\varphi'(\tilde{y})$, $\hat{\varphi}\equiv 0$ and $\hat{f}=-(Q\tilde{y}-\hat{q})$.
- For $\hat{\kappa}_{\mathrm{ref}}\equiv\kappa$, $\hat{\chi}_{\mathrm{ref}}\equiv 0$, we obtain $\hat{A}_{\mathrm{ref}}=A$. For $\hat{\kappa}_{\mathrm{ref}}\equiv 1$ and $\hat{\chi}_{\mathrm{ref}}\equiv 0$ we get the $H_0^1(\Omega)$-norm inducing operator. Changing to $\hat{\chi}_{\mathrm{ref}}\equiv 1$, the $H^1(\Omega)$-norm is induced.

We use the FE discretization described in subsection 5.1 with the triangulation $\mathscr{T}$ and assume that the coefficient functions $\hat{\kappa}$) and $\hat{\kappa}_{\mathrm{ref}}$ are piecewise constant on the triangles. This is done to simplify the explanations and results in this section. Quadrilateral elements, cf. [1], higher order FE functions, or not piecewise constant coefficient functions could be included quite simply, but would result in more complicated formulas or distinctions of cases.

With $v\in Y$ we get

$$\begin{aligned}
&(w-\mathrm{w},v)_{\hat{A}_{\mathrm{ref}}} = \langle r-\hat{A}_{\mathrm{ref}}\mathrm{w},v\rangle_{Y^*,Y} =\\
&\sum_{T\in\mathscr{T}}\Big(\int_T \hat{\kappa}\nabla\tilde{y}\cdot\nabla v+\hat{\chi}\tilde{y}v+\hat{\varphi}(\tilde{y})v-\hat{f}v-\hat{\kappa}_{\mathrm{ref}}\nabla\mathrm{w}\cdot\nabla v-\hat{\chi}_{\mathrm{ref}}\mathrm{w}v\,\mathrm{d}x\Big) =\\
&\sum_{T\in\mathscr{T}}\Big(\int_T -\mathrm{div}(\hat{\kappa}\nabla\tilde{y})v+\mathrm{div}(\hat{\kappa}_{\mathrm{ref}}\nabla\mathrm{w})v+\big(\hat{\chi}\tilde{y}+\hat{\varphi}(\tilde{y})-\hat{f}-\hat{\chi}_{\mathrm{ref}}\mathrm{w}\big)v\,\mathrm{d}x\\
&\qquad+\int_{\partial T}\big(\hat{\kappa}\nabla\tilde{y}-\hat{\kappa}_{\mathrm{ref}}\nabla\mathrm{w}\big)\cdot\mathfrak{n}_T\; v\,\mathrm{d}S\Big),
\end{aligned}$$

where $\mathfrak{n}_T$ is the outer unit normal of the triangle T. Since $\hat{\kappa}$ and $\hat{\kappa}_{\mathrm{ref}}$ are piecewise constant on the triangles and we use piecewise linear ansatz functions, $\mathrm{div}(\hat{\kappa}_{\mathrm{ref}}\nabla\mathrm{w})\equiv 0\equiv\mathrm{div}(\hat{\kappa}\nabla\tilde{y})$ holds on each element T. Therefore, using $(w-\mathrm{w},\mathrm{v})_{\hat{A}_{\mathrm{ref}}}=0$, we get

$$\begin{aligned}
&(w-\mathrm{w},v)_{\hat{A}_{\mathrm{ref}}} = (w-\mathrm{w},v-\mathrm{v})_{\hat{A}_{\mathrm{ref}}} =\\
&\sum_{T\in\mathscr{T}}\Big(\int_T\big(\hat{\chi}\tilde{y}+\hat{\varphi}(\tilde{y})-\hat{f}-\hat{\chi}_{\mathrm{ref}}\mathrm{w}\big)(v-\mathrm{v})\,\mathrm{d}x+\int_{\partial T}\big(\hat{\kappa}\tfrac{\partial}{\partial\mathfrak{n}_T}\tilde{y}-\hat{\kappa}_{\mathrm{ref}}\tfrac{\partial}{\partial\mathfrak{n}_T}\mathrm{w}\big)(v-\mathrm{v})\,\mathrm{d}S\Big)
\end{aligned}$$

for arbitrary $\mathrm{v}\in\mathrm{Y}$. Since $\tilde{y}$ is bounded and continuous on $\overline{\Omega}$ and $\hat{\varphi}:\mathbb{R}\to\mathbb{R}$ is continuous, $\hat{\varphi}(\tilde{y})$ belongs to $L^2(\Omega)$. Due to $\hat{\chi},\hat{\chi}_{\mathrm{ref}}\in L^\infty(\Omega)$, we also have $\hat{\chi}\tilde{y},\hat{\chi}_{\mathrm{ref}}\mathrm{w}\in L^2(\Omega)$. The integrals over the triangle boundaries ∂T are considered edge-wise: If an edge E belongs to $\partial\Omega$, this part of the integral vanishes. Interior edges also appear in the integral over the boundary of a neighboring triangle. Thus, the sum over all triangle boundary integrals can be collected to integrals over all interior edges $E\in\mathscr{E}^0$, denoting by $\mathscr{E}$ the set of all edges and by $\mathscr{E}^0$ the set of all interior edges. There, the normal jumps involving the discrete solutions $\tilde{y}$ and w appear:

$$[\![\phi]\!]_E(x) := \lim_{t\to 0+}\phi(x+t\mathfrak{n}_E)-\lim_{t\to 0+}\phi(x-t\mathfrak{n}_E),$$

where $\mathfrak{n}_E$ is a unit normal vector corresponding to E, cf. [20, sec. 1.1].

We estimate

$$
\begin{aligned}
(w-\mathrm{w},v)_{\hat{A}_{\mathrm{ref}}} \leq & \sum_{T\in\mathscr{T}} \|\hat{\chi}\tilde{\mathrm{y}}+\hat{\varphi}(\tilde{\mathrm{y}})-\hat{f}-\hat{\chi}_{\mathrm{ref}}\mathrm{w}\|_{L^2(T)}\|v-\mathrm{v}\|_{L^2(T)} \\
& + \sum_{E\in\mathscr{E}^0} \|[\![(\hat{\kappa}\nabla\tilde{\mathrm{y}}-\hat{\kappa}_{\mathrm{ref}}\nabla\mathrm{w})\cdot\mathfrak{n}_E]\!]_E\|_{L^2(E)}\|v-\mathrm{v}\|_{L^2(E)}
\end{aligned} \tag{5.21}
$$

If we insert the Clément interpolant $\mathrm{v}\in\mathrm{Y}$ of v and use that $(w-\mathrm{w},v)_{\hat{A}_{\mathrm{ref}}} \leq c\cdot\eta(\mathrm{w})\cdot\|v\|_{\hat{A}_{\mathrm{ref}}}$ for all $v\in Y$ yields $\|w-\mathrm{w}\|_{\hat{A}_{\mathrm{ref}}}\leq c\cdot\eta(\mathrm{w})$, a similar estimation as in [20, sec. 1.2] or [1, sec. 2.2] can be performed, giving

$$
\|w-\mathrm{w}\|_{\hat{A}_{\mathrm{ref}}} \leq c_{\mathscr{T}}c_{\hat{A}_{\mathrm{ref}}}\Big(\sum_{T\in\mathscr{T}}\eta_T(\tilde{\mathrm{y}})^2+\sum_{E\in\mathscr{E}^0}\eta_E(\tilde{\mathrm{y}})^2\Big)^{1/2}, \quad \text{where} \tag{5.22}
$$

$$
\eta_T(\tilde{\mathrm{y}}) := h_T\|\hat{\chi}\tilde{\mathrm{y}}+\hat{\varphi}(\tilde{\mathrm{y}})-\hat{f}-\hat{\chi}_{\mathrm{ref}}\mathrm{w}\|_{L^2(T)}, \tag{5.23}
$$

$$
\eta_E(\tilde{\mathrm{y}}) := h_E^{1/2}\|[\![(\hat{\kappa}\nabla\tilde{\mathrm{y}}-\hat{\kappa}_{\mathrm{ref}}\nabla\mathrm{w})\cdot\mathfrak{n}_E]\!]_E\|_{L^2(E)}. \tag{5.24}
$$

The constant $c_{\mathscr{T}}>0$ depends only on the smallest angle in the triangulation $\mathscr{T}$ and the coercivity constant $c_{\hat{A}_{\mathrm{ref}}}>0$ is chosen such that $\|y\|_{H^1(\Omega)}\leq c_{\hat{A}_{\mathrm{ref}}}\|y\|_{\hat{A}_{\mathrm{ref}}}$ holds for all $y\in Y$. The diameters of the triangles and edges are denoted by h_T and h_E, respectively.

The overall error estimator looks as follows:

THEOREM 5.4. *Let $\tilde{\mathrm{y}}\in\mathrm{Y}$ (linear finite element space on the polygonal domain $\Omega\subset\mathbb{R}^2$) be given and let $y\in Y$ be the exact solution of (5.9), where the respective operators are defined as in (5.20). Furthermore, let $\hat{\kappa}$ and $\hat{\kappa}_{\mathrm{ref}}$ be piecewise constant and define* w *by (5.16) and (5.12). Then,*

$$
\|\tilde{\mathrm{y}}-y\|^2_{\hat{A}_{\mathrm{ref}}} \leq \Lambda^2\|\mathrm{w}\|^2_{\hat{A}_{\mathrm{ref}}}+\Lambda^2c^2_{\mathscr{T}}c^2_{\hat{A}_{\mathrm{ref}}}\Big(\sum_{T\in\mathscr{T}}\eta_T(\tilde{\mathrm{y}})^2+\sum_{E\in\mathscr{E}^0}\eta_E(\tilde{\mathrm{y}})^2\Big) \tag{5.25}
$$

holds with Λ from (5.10), $c_{\mathscr{T}},c_{A_{\mathrm{ref}}}$ from (5.22), and $\eta_T(\tilde{\mathrm{y}}),\eta_E(\tilde{\mathrm{y}})$ as in (5.23), (5.24).

Proof. Combining (5.17) and (5.22) yields the desired result. ∎

In practice, we obtain always a negligible algebraic error $\|\mathrm{w}\|_{\hat{A}_{\mathrm{ref}}}$ by applying Newton's method to the discrete equation. To adaptively solve (5.9) with the operators (5.20), we assign half of the edge error $\eta_E(\tilde{\mathrm{y}})$ to each of the two neighboring triangles. Based on that, we mark all triangles with the largest error contributions which constitute a certain amount $\vartheta_\eta\in(0,1)$ of the total error, e.g., $\vartheta_\eta=30\%$ in our implementation, see [7, sec. 7.1], a so-called Dörfler strategy [6]. These triangles are refined regularly, i.e., divided into four triangles of the same shape. To avoid hanging nodes, additional triangles have to be divided into two new ones possibly.

6. Implementation and numerical results. The inexact trust-region method (Algorithm 1) is implemented in MATLAB to solve different instances of the model problem presented in section 2. In order to solve the arising PDEs, the adaptive solution technique described in section 5 is used. The objective function evaluation and gradient error are estimated as derived in section 4. For this purpose, we consider the following concrete setups of the model problem (2.3):

- The domain $\Omega:=(-1,1)^2\setminus(-1,0]^2\subset\mathbb{R}^2$ is the polygonal L-shaped domain. The initial FE mesh contains 113 nodes.
- The coefficient function is $\kappa\equiv 1$ and the reference coefficients are $\hat{\kappa}_{\mathrm{ref}}\equiv 1$ and $\hat{\chi}_{\mathrm{ref}}\equiv 0$ so that $\|\cdot\|_{A_{\mathrm{ref}}}\equiv\|\cdot\|_{H^1_0(\Omega)}$ and the constant $c_{A_{\mathrm{ref}}}$ introduced in (5.22) comes from the Poincaré inequality $\|\cdot\|_{H^1(\Omega)}\leq c_{A_{\mathrm{ref}}}\|\cdot\|_{H^1_0(\Omega)}$. In (5.10) we have $\Lambda=1$.

- $U = L^2(\Omega)$, i.e., $\Omega_u = \Omega$, $D \equiv I : L^2(\Omega) \to L^2(\Omega)$, $f \equiv 0$, and $\varphi(t) := t^3$.
- The observation space is $H = L^2(\Omega)$ and $Q \equiv \iota : H_0^1(\Omega) \hookrightarrow\hookrightarrow L^2(\Omega)$.
- The desired state $\hat{q}(\cdot) \equiv \hat{q} \in L^2(\Omega)$ is $\hat{q}(x) = 1$ (first setup) and $\hat{q}(x) = 9(x_1^3 - x_1)(x_2^3 - x_2)$ (second setup), a polynomial which fulfills the zero boundary conditions on Ω and is very smooth around the corner point $x = 0$.
- We take $\gamma = 10^{-3}$ and $U_{\text{ad}} := \{u \in L^2(\Omega) : u(x) \leq 14 \text{ for a.e. } x \in \Omega\}$, where the upper bound is chosen such that it becomes active, but the optimal state is in the order of magnitude of the desired state.
- We initialize the algorithm with $u^0 = 0$.

The constants required for error estimation are set as follows: In Theorem 4.1 and in Lemma 4.2, we choose the Sobolev constant $c_{\tilde{r}} = c_p = c_4 = 0.5$ because $c_2 \approx 0.3$ can be estimated numerically. In (5.22) we choose $c_{\mathscr{T}} c_{A_{\text{ref}}} = 1$. Even if the real constants are underestimated by these choices, the algorithm still works because we need to know the error only up to a fixed, possibly unknown multiplicative constant. In fact, an unrealistic choice of constants can be compensated by the choice of the error functions in the trust-region algorithm.

Implementation details. Algorithm 1 is implemented such that it is suitable for optimal control applications. Therefore, the objective function and gradient evaluations accept initializations for the state and the adjoint state, a refined version of which is returned and used for further computations. All relevant functions, such as the control, the state, and the adjoint state are kept on matched grids to facilitate the computation. We keep all grid refinements stemming from the gradient computation and the projected linesearch for further iterations, but discard the ones caused by objective function evaluations, which sometimes requires high accuracy. Keeping a fine FE space for further iterations would make the algorithm slow. To find a possibly better step than the generalized Cauchy step found by the projected Armijo linesearch with suitably refined projection, we apply a semismooth Newton method [17, 18, 13, 19] with the generalized Cauchy point as initial iterate. As it might be difficult to handle the control constraints and the trust-region constraint simultanously, we replace the latter in (3.3) by a quadratic regularization. The resulting search direction is projected onto $U_{\text{ad}} - u^k$ and possibly scaled such that it satisfies the constraints of (3.3). For the application of semismooth Newton in function space we refer to [14, chap. 2] and [19] and for details of an implementation for a discretized problem to [11].

To make sure that all required error bounds in Algorithm 1 are satisfied, we use the strategy proposed in subsection 3.3. The parameter settings for the trust-region algorithm as well as the inexact projection and the projected linesearch are listed in Table 1. Experimentation for

Algorithm 1, see subsections 3.1 and 3.3:

τ	$\tau = \frac{1}{\gamma} = 10^3$
$\tilde{\mathfrak{c}}$, $\mathfrak{e}$	$\tilde{\mathfrak{c}}_{\text{c}} = 200$, $\tilde{\mathfrak{c}}_{\text{g}} = 200$, $\tilde{\mathfrak{c}}_{\text{p}} = 0$ (exact projection), $\tilde{\mathfrak{c}}_{\text{o}} = 10^8$, $\mathfrak{e}_{\text{o}} = 1.1$
$(\mathfrak{r}_k)_{k \in \mathbb{N}_0}$	$\mathfrak{r}_k = \frac{1000}{k+1}$
Δ	$\Delta_{\max} = 10^4$, $\Delta_0 = 1$
η_i, ν_i	$\eta_1 = 0.3$, $\eta_2 = 0.7$, $\eta_3 = 0.2$, $\quad \nu_1 = 0.5$, $\nu_2 = 1.0$, $\nu_3 = 2.0$

Inexact projection and projected linesearch, see subsection 3.3:

$\mathfrak{c}$	$\mathfrak{c}_{\text{i}} = 0.5$, $\mathfrak{c}_{\text{f}} = 0.5$, $\mathfrak{c}_{\text{e}} = 10^{-2}$, $\mathfrak{c}_{\text{a}} = 10^{-3}$, $\mathfrak{c}_{\text{d}} = 10^{-2}$
$\tilde{\mathfrak{c}}_{\text{l1}}$, $\tilde{\mathfrak{c}}_{\text{l2}}$	$\tilde{\mathfrak{c}}_{\text{l1}} = \frac{3}{7}$ ($c_{\text{p}} = 1$, $\tilde{\mathfrak{c}}_{\text{l1}} = 0.7$), $\quad \tilde{\mathfrak{c}}_{\text{l2}} = \frac{1}{1 - c_{\text{p}} \tilde{\mathfrak{c}}_{\text{l1}}} = \frac{7}{4}$, see Remark 3.9

TABLE 1
Parameters used in Algorithm 1

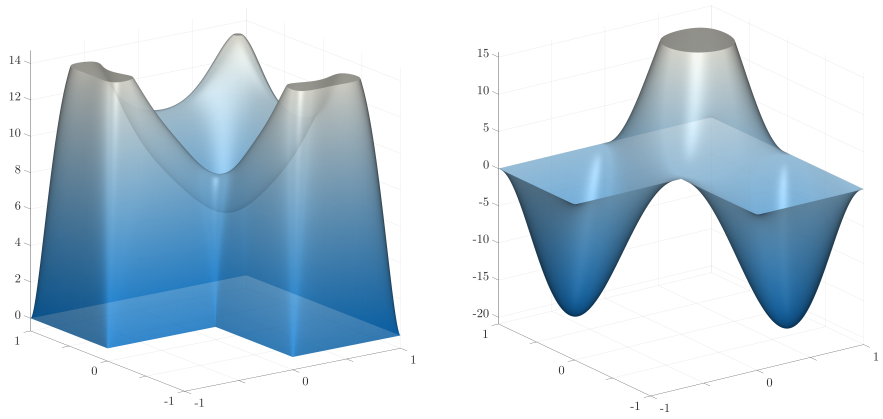

FIG. 1. *Optimal controls for the* $\hat{q} \equiv 1$ *(left) and* $\hat{q}(x) = 9(x_1^3 - x_1)(x_2^3 - x_2)$ *(right).*

the parameter choices was necessary in case of the error bound parameters $\tilde{\mathfrak{c}}_c > 0$, $\tilde{\mathfrak{c}}_g > 0$, and $\tilde{\mathfrak{c}}_o > 0$ because they reflect how good the applied error estimators are or compensate unknown constants. They have been chosen such that no unsuccessful iterations occur and the grid refinement happens not too fast. In particular, the high value chosen for $\tilde{\mathfrak{c}}_o$ comes also from the fact that we overestimate the error given in Proposition 4.3, because we measure the state error in the $H_0^1(\Omega)$-norm although the $L^2(\Omega)$-error would be sufficient.

In the adaptive solution of the PDEs and the adaptive computation of the projection, we stop the refinement if $2 \cdot 10^5$ FE nodes would be exceeded so that the algorithm solves the discretized problem on the finest mesh at the end. We stop the algorithm if $\chi_k(0) < 10^{-4}$. Note that in the unconstrained case this would correspond to having $\|\nabla m_k(0)\| < 10^{-7}$ because $\tau = 10^3$ is chosen, cf. (3.8).

Results. The computed optimal controls are depicted in Figure 1. The different active sets and (lacking) smoothness around $x = 0$ can be recognized. Details of the FE meshes obtained in the final iteration of the algorithm are shown in Figure 2. Around the corner $x = 0$, the mesh is refined locally in the first setup due to the adaptive solution of the PDE. This does not happen in the second setup, where the solution is smooth around this point. Furthermore, the boundaries of the respective active sets of the controls are resolved by the meshes due to the refined projection.

The convergence for the first setup is shown in Figure 3. The number of FE nodes increases until it reaches the allowed upper bound. The criticality measure decreases in general until the desired tolerance is reached. But by refining the mesh the inexactness of the computed criticality measure is recognized sometimes. Then the computed criticality measure on the refined mesh is larger than the one on the coarser grid. We have to mention that it is not worth counting iterations in this setting. Typically, the first iterations run very fast within some seconds. For instance, the first 16 iterations needed to decrease the computed optimality measure from 23.2 to $3.09 \cdot 10^{-2}$ take about 19 seconds in total on our machine. Only the last iterations required to obtain the desired high accuracy last increasingly longer so that the total computing time is around 22 minutes. The last three iterations, during which the computed criticality measure is decreased from $1.23 \cdot 10^{-3}$ to $0.98 \cdot 10^{-4}$, take approximately 13 minutes. Thus, for reducing the criticality measure to approximately 10^{-3}, which due to $\tau = 10^3$ corresponds to a gradient norm of 10^{-6} in the unconstrained case, takes approximately 9 minutes. Concerning these runtimes, it should be mentioned that our MATLAB implementation was not explicitly optimized for speed.

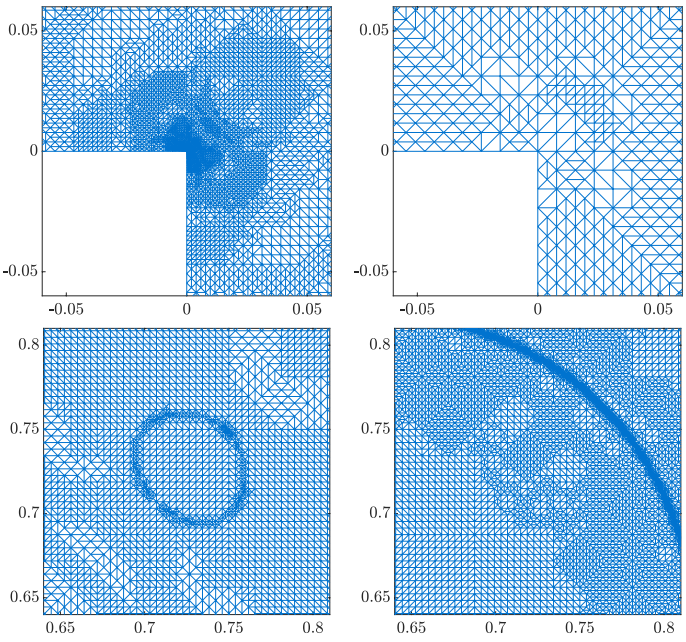

FIG. 2. *Details of the final meshes for* $\hat{q} \equiv 1$ *(left) and* $\hat{q}(x) = 9(x_1^3 - x_1)(x_2^3 - x_2)$ *(right).*

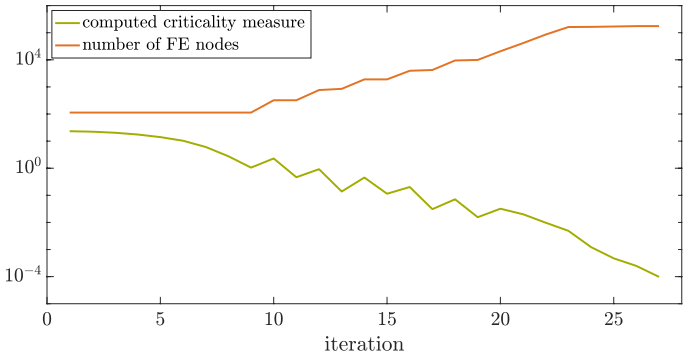

FIG. 3. *Convergence and refinement plot for* $\hat{q} \equiv 1$.

Appendix A. Convergence proof for Algorithm 1. We apply the following two lemmas to prove Theorem 3.3:

LEMMA A.1. *Let Assumptions 3.1 and 3.2 hold and let the sequence of inexact criticality measures (as defined in (3.8)) generated by Algorithm 1 satisfy*

$$\chi_k(0) \geq \varepsilon > 0 \quad \text{for all } k \geq K_1 \in \mathbb{N}_0 \tag{A.1}$$

for some fixed $\varepsilon > 0$. *Then,* $\lim_{k\to\infty} \Delta_k = 0$ *holds for the sequence of corresponding trust-region radii.*

Proof. First observe that the termination criterion "$\chi_k(0) = 0$" in the algorithm is not met for *any* $k \in \mathbb{N}_0$ by assumption, because $\chi_k(0) = 0$ for some $k \in \mathbb{N}_0$ would yield $\chi_\ell(0) = 0$ for all $\ell \geq k$, which contradicts (A.1). Moreover, $\text{pred}_k > 0$ holds for all $k \in \mathbb{N}_0$ due to (3.12), the positivity property of ρ_{t1} and ρ_{t2}, and $\chi_k(0) > 0$, $\Delta_k > 0$. Due to $\lim_{k\to\infty} \tau_k = 0$ and $\rho_{\text{r}}(t) \leq t$

for small enough t, it holds that

$$\rho_{\mathrm{r}}(\eta_3 \min\{\mathrm{pred}_k, \mathfrak{r}_k\}) \leq \eta_3 \mathrm{pred}_k \quad \text{for all } k \geq K_2 \in \mathbb{N}_0$$

with some $K_2 \geq K_1$. By (3.11) we thus get

$$|\mathrm{ared}_k - \mathrm{cred}_k| \leq \eta_3 \mathrm{pred}_k \quad \text{for all } k \geq K_2.$$

This implies

$$\mathrm{ared}_k = \mathrm{cred}_k + \mathrm{ared}_k - \mathrm{cred}_k \geq \mathrm{cred}_k - \eta_3 \mathrm{pred}_k = \big(\tfrac{\mathrm{cred}_k}{\mathrm{pred}_k} - \eta_3\big)\mathrm{pred}_k \tag{A.2}$$

for all $k \geq K_2$. This is well-defined due to $\mathrm{pred}_k > 0$.
Now we show that $\sum_{k=0}^{\infty} \Delta_k < \infty$ follows from (A.1). For all unsuccessful steps $k \in \mathscr{I}_{\mathrm{u}} \subset \mathbb{N}_0$ we have $\Delta_{k+1} \leq \nu_1 \Delta_k$ with the parameter $\nu_1 \in (0,1)$. Thus, if there are only finitely many (very) successful steps, the sequence $(\Delta_k)_{k\in\mathbb{N}_0}$ is summable since then $\Delta_k \leq {\nu_1}^{k-K_3}\Delta_{K_3}$ for all $k \geq K_3$ for some $K_3 \in \mathbb{N}_0$. In the following, we consider the case of infinitely many (very) successful steps. For a (very) successful step $k \in \mathscr{I}_{\mathrm{s}} = \mathbb{N}_0 \setminus \mathscr{I}_{\mathrm{u}}$, i.e.,

$$\tfrac{\mathrm{cred}_k}{\mathrm{pred}_k} \geq \eta_1 \text{ and } \Delta_{k+1} \leq \min\{\nu_3\Delta_k, \Delta_{\max}\} \leq \nu_3\Delta_k,$$

we deduce from (A.2) and (3.12):

$$\begin{aligned} \mathrm{ared}_k &\geq \big(\tfrac{\mathrm{cred}_k}{\mathrm{pred}_k} - \eta_3\big)\mathrm{pred}_k \geq (\eta_1 - \eta_3)\mathrm{pred}_k \\ &\geq (\eta_1 - \eta_3)\cdot \rho_{\mathrm{t}1}(\chi_k(0)) \cdot \min\{\rho_{\mathrm{t}2}(\chi_k(0)), \Delta_k\} \\ &\geq (\eta_1 - \eta_3)\cdot \rho_{\mathrm{t}1}(\varepsilon) \cdot \min\{\rho_{\mathrm{t}2}(\varepsilon), \Delta_k\} > 0, \end{aligned} \tag{A.3}$$

for all $k \in \mathscr{I}_{\mathrm{s}}$, $k \geq K_2$, where we used that $\rho_{\mathrm{t}1}$ and $\rho_{\mathrm{t}2}$ are increasing. Since by Assumption 3.1, the sequence $(\hat{J}(u^k))_{k\in\mathscr{I}_{\mathrm{s}}}$ is bounded from below, we get

$$0 \leq \sum_{k\in\mathscr{I}_{\mathrm{s}}, k\geq K_2} \mathrm{ared}_k = \sum_{k\in\mathscr{I}_{\mathrm{s}}, k\geq K_2} \big(\hat{J}(u^k) - \hat{J}(u^{k+1})\big) = \sum_{k=K_2}^{\infty} \big(\hat{J}(u^k) - \hat{J}(u^{k+1})\big) < \infty \tag{A.4}$$

using that $u^{k+1} = u^k + s^k$ in the case of a (very) successful step and $u^{k+1} = u^k$ in the case of an unsuccessful step. Due to $\eta_1 > \eta_3$, $\rho_{\mathrm{t}1}(\varepsilon) > 0$, $\rho_{\mathrm{t}2}(\varepsilon) > 0$ and $\lim_{\mathscr{I}_{\mathrm{s}}\ni k\to\infty} \mathrm{ared}_k = 0$ (by (A.4)), it follows from (A.3) that $\Delta_k \leq \frac{\mathrm{ared}_k}{(\eta_1-\eta_3)\cdot\rho_{\mathrm{t}1}(\varepsilon)}$ for all $k \in \mathscr{I}_{\mathrm{s}}$, $k \geq K_4$, with some sufficiently large $K_4 \in \mathbb{N}_0$, $K_4 \geq K_2$, and thus, by (A.4),

$$0 \leq \sum_{k\in\mathscr{I}_{\mathrm{s}}} \Delta_k < \infty. \tag{A.5}$$

Now we consider two (very) successful steps $\tilde{k}, \hat{k} \in \mathscr{I}_{\mathrm{s}}$, $\hat{k} \geq \tilde{k}+2$ with only unsuccessful steps $k \in \{\tilde{k}+1, \dots, \hat{k}-1\}$ in between. Hence, we have $\Delta_k \leq \nu_3 {\nu_1}^{k-\tilde{k}-1}\Delta_{\tilde{k}}$ for $\tilde{k}+1 \leq k \leq \hat{k}-1$ and thus (using the geometric series with $\nu_1 \in (0,1)$)

$$\Sigma(\tilde{k}) := \sum_{k=\tilde{k}}^{\hat{k}-1} \Delta_k \leq \Delta_{\tilde{k}}\Big(1 + \nu_3 \sum_{\ell=0}^{\hat{k}-\tilde{k}-2} {\nu_1}^{\ell}\Big) \leq \Delta_{\tilde{k}}\Big(1 + \tfrac{\nu_3}{1-\nu_1}\Big).$$

Additionally, for $\tilde{k} \in \mathscr{I}_{\mathrm{s}}$ such that $\tilde{k}+1 \in \mathscr{I}_{\mathrm{s}}$ we set $\Sigma(\tilde{k}) = \Delta_{\tilde{k}}$ and in the case $0 \notin \mathscr{I}_{\mathrm{s}}$ we set and estimate

$$\Sigma(0) := \sum_{k=0}^{\hat{k}-1} \Delta_k \leq \Delta_0\Big(\sum_{\ell=0}^{\hat{k}-1} {\nu_1}^{\ell}\Big) \leq \Delta_0 \cdot \tfrac{1}{1-\nu_1},$$

where $\hat{k} = \min \mathscr{I}_{\mathrm{s}}$. Therefore,

$$0 \leq \sum_{k=0}^{\infty} \Delta_k = \sum_{\tilde{k} \in \mathscr{I}_{\mathrm{s}} \cup \{0\}} \Sigma(\tilde{k}) \leq \left(1 + \tfrac{\nu_3}{1-\nu_1}\right) \cdot \sum_{\tilde{k} \in \mathscr{I}_{\mathrm{s}} \cup \{0\}} \Delta_{\tilde{k}} < \infty$$

follows with $\Sigma(\tilde{k}) \leq \Delta_{\tilde{k}}\left(1 + \frac{\nu_3}{1-\nu_1}\right)$ for all $\tilde{k} \in \mathscr{I}_{\mathrm{s}} \cup \{0\}$ (using $\nu_3 \geq 1$) and (A.5). We see that $(\Delta_k)_{k \in \mathbb{N}_0}$ is summable and thus $\lim_{k\to\infty} \Delta_k = 0$. ∎

LEMMA A.2. *Under Assumptions 3.1 and 3.2*

$$\liminf_{k\to\infty} \chi_k(0) = 0 \tag{A.6}$$

holds true for every sequence generated by Algorithm 1, where the inexact criticality measure χ_k is defined as in (3.8).

Proof. For a proof by contradiction, assume that (A.6) is false, giving that (A.1) is true for some fixed $\varepsilon > 0$. By Lemma A.1 we have that $\lim_{k\to\infty} \Delta_k = 0$ and thus $\lim_{k\to\infty} \|s^k\|_U = 0$. In analogy to (A.2), we can estimate

$$\mathrm{cred}_k \geq \mathrm{ared}_k - |\mathrm{ared}_k - \mathrm{cred}_k| \geq \mathrm{ared}_k - \eta_3 \mathrm{pred}_k \tag{A.7}$$

for all $k \geq K_2 \geq K_1$. As in (A.3) we infer

$$\mathrm{pred}_k \geq \rho_{\mathrm{t}1}(\varepsilon) \cdot \Delta_k$$

for all $k \geq K_5$ with some sufficiently large $K_5 \in \mathbb{N}_0$, $K_5 \geq K_2$, due to (3.12) and $\lim_{k\to\infty} \Delta_k = 0$, i.e., $\Delta_k \leq \rho_{\mathrm{t}2}(\varepsilon)$ for all $k \geq K_5$ because $\rho_{\mathrm{t}2}(\varepsilon) > 0$. Thus,

$$|o(\Delta_k)| \leq (1 - \eta_3 - \eta_2)\mathrm{pred}_k \tag{A.8}$$

for all $k \geq K_6$ with $K_6 \geq K_5 \geq K_2$ sufficiently large, since $(1 - \eta_3 - \eta_2) > 0$. Using this and the bounds indicated below, we estimate for $k \geq K_6$ and using $s_k \to 0$:

$$\begin{aligned}
\mathrm{cred}_k &\overset{(A.7)}{\geq} \mathrm{ared}_k - \eta_3 \mathrm{pred}_k \overset{(3.2)}{\geq} -(\nabla \hat{J}(u^k), s^k)_U - \eta_3 \mathrm{pred}_k - |o(\|s^k\|)| \\
&\overset{(3.13)}{\geq} \mathrm{pred}_k + (\nabla m_k(0), s^k)_U - (\nabla J(u^k), s^k)_U - \eta_3 \mathrm{pred}_k - |o(\|s^k\|)| \\
&\geq (1 - \eta_3)\mathrm{pred}_k - \|\nabla m_k(0) - \nabla \hat{J}(u^k)\|_U \|s^k\|_U - |o(\|s^k\|)| \\
&\overset{\substack{(3.9),\\ \|s^k\|_U \leq \Delta_k}}{\geq} (1 - \eta_3)\mathrm{pred}_k - \rho_{\mathrm{g}}(\Delta_k)\Delta_k - |o(\Delta_k)| \overset{\substack{\rho_{\mathrm{g}}(t)\to 0\\ (t\to 0^+)}}{\geq} (1 - \eta_3)\mathrm{pred}_k - |o(\Delta_k)| \\
&\overset{(A.8)}{\geq} (1 - \eta_3)\mathrm{pred}_k - (1 - \eta_3 - \eta_2)\mathrm{pred}_k = \eta_2 \mathrm{pred}_k
\end{aligned}$$

Note that $u^k \in \{u \in U_{\mathrm{ad}} : \hat{J}(u) \leq \hat{J}(u^{K_2})\}$ holds for all $k \geq K_2$ due to (A.3) so that (3.2) is applicable. In fact, the objective function values $(\hat{J}(u^k))_{k \geq K_2}$ are non-increasing since (A.3) holds for (very) successful steps and the function values do not change for unsuccessful steps. Using $\mathrm{pred}_k > 0$ as in the proof of Lemma A.1, it follows that $\frac{\mathrm{cred}_k}{\mathrm{pred}_k} \geq \eta_2$ for all $k \geq K_6$, i.e., all steps $k \geq K_6$ are successful giving $\Delta_{k+1} \geq \min\{\nu_2 \Delta_k, \Delta_{\max}\} \geq \Delta_k > 0$ due to $\nu_2 \geq 1$. This contradicts $\lim_{k\to\infty} \Delta_k = 0$, proving (A.6). ∎

Using Lemma A.2, the proof of Theorem 3.3 is very short:

Proof of Theorem 3.3. Due to (3.10) we have

$$\chi(u^k) \leq \chi_k(0) + |\chi_k(0) - \chi(u^k)| \leq \chi_k(0) + \rho_{\mathrm{c}}(\chi_k(0)).$$

This is also true if the algorithm is stopped due to $\chi_k(0) = 0$, because then $\chi(u^k) = 0$ follows from (3.10). Therefore, $\chi(u^\ell) = \chi(u^k) = 0 = \chi_\ell(0)$ for all $\ell \geq k$. The bound on $\chi(u^k)$, $\lim_{t\to 0^+} \rho_c(t) = 0$, and $\rho_c(0) = 0$ show $0 \leq \liminf_{k\to\infty} \chi(u^k) \leq \liminf_{k\to\infty} \chi_k(0) + \rho_c(\chi_k(0)) = 0$. ∎

REFERENCES

[1] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Pure and Applied Mathematics, John Wiley & Sons, New York, 2000, https://doi.org/10.1002/9781118032824.

[2] R. Becker and R. Rannacher, *An optimal control approach to a posteriori error estimation in finite element methods*, Acta Numer., 10 (2001), pp. 1–102, https://doi.org/10.1017/S0962492901000010.

[3] A. Bespalov, C. E. Powell, and D. Silvester, *Energy norm a posteriori error estimation for parametric operator equations*, SIAM J. Sci. Comput., 36 (2014), pp. A339 – A363, https://doi.org/10.1137/130916849.

[4] A. Bespalov and D. Silvester, *Efficient adaptive stochastic Galerkin methods for parametric operator equations*, SIAM J. Sci. Comput., 38 (2016), pp. A2118 – A2140, https://doi.org/10.1137/15M1027048.

[5] A. Conn, N. Gould, and P. Toint, *Trust-Region Methods*, SIAM and MOS, Philadelphia, 2000, https://doi.org/10.1137/1.9780898719857.

[6] W. Dörfler, *A Convergent Adaptive Algorithm for Poisson's Equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124, https://doi.org/10.1137/0733054.

[7] M. Eigel, C. J. Gittelson, C. Schwab, and E. Zander, *Adaptive stochastic Galerkin FEM*, Comput. Methods Appl. Mech. Engrg., 270 (2014), pp. 247–269, https://doi.org/10.1016/j.cma.2013.11.015.

[8] M. Eigel, C. J. Gittelson, C. Schwab, and E. Zander, *A convergent adaptive stochastic Galerkin finite element method with quasi-optimal spatial meshes*, ESAIM Math. Model. Numer. Anal., 49 (2015), pp. 1367–1398, https://doi.org/10.1051/m2an/2015017.

[9] M. Eigel, M. Pfeffer, and R. Schneider, *Adaptive stochastic Galerkin FEM with hierarchical tensor representations*, Numer. Math., 136 (2017), pp. 765–803, https://doi.org/10.1007/s00211-016-0850-x.

[10] S. Garreis, *Optimal Control under Uncertainty: Theory and Numerical Solution with Low-Rank Tensors*, PhD thesis, Technische Universität München, 2019.

[11] S. Garreis and M. Ulbrich, *Constrained Optimization with Low-Rank Tensors and Applications to Parametric Problems with PDEs*, SIAM J. Sci. Comput., 39 (2017), pp. A25–A54, https://doi.org/10.1137/16M1057607.

[12] S. Garreis and M. Ulbrich, *A Fully Adaptive Algorithm for PDE-Constrained Optimal Control with and without Uncertainty,, Part 2: The Stochastic Case and Low-Rank Tensor Methods*, Preprint, submitted, Technical University of Munich, 2019.

[13] M. Hintermüller, K. Ito, and K. Kunisch, *The Primal-Dual Active Set Strategy as a Semismooth Newton Method*, SIAM J. Optim., 13 (2002), pp. 865–888, https://doi.org/10.1137/S1052623401383558.

[14] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Springer, New York, 2009, https://doi.org/10.1007/978-1-4020-8839-1.

[15] D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders, *A Trust-Region Algorithm with Adaptive Stochastic Collocation for PDE Optimization under Uncertainty*, SIAM J. Sci. Comput., 35 (2013), pp. A1847–A1879, https://doi.org/10.1137/120892362.

[16] D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders, *Inexact Objective Function Evaluations in a Trust-Region Algorithm for PDE-Constrained Optimization under Uncertainty*, SIAM J. Sci. Comput., 36 (2014), pp. A3011–A3029, https://doi.org/10.1137/140955665.

[17] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367, https://doi.org/10.1007/BF01581275.

[18] M. Ulbrich, *Semismooth Newton Methods for Operator Equations in Function Spaces*, SIAM J. Optim., 13 (2002), pp. 805–841, https://doi.org/10.1137/S1052623400371569.

[19] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, vol. 11 of MOS-SIAM Ser. Optim., SIAM, Philadelphia, 2011, https://doi.org/10.1137/1.9781611970692.

[20] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Chichester and Stuttgart, 1996.

[21] E. Zeidler, *Nonlinear Functional Analysis and its Applications*, vol. II/B: Nonlinear Monotone Operators, Springer, New York, 1990, https://doi.org/10.1007/978-1-4612-0981-2.

[22] J. C. Ziems, *Adaptive Multilevel Inexact SQP Methods for PDE-Constrained Optimization with Control Constraints*, SIAM J. Optim., 23 (2013), pp. 1257–1283, https://doi.org/10.1137/110848645.

[23] J. C. Ziems and S. Ulbrich, *Adaptive Multilevel Inexact SQP Methods for PDE-Constrained Optimization*, SIAM J. Optim., 21 (2011), pp. 1–40, https://doi.org/10.1137/080743160.